BTI and Leakage Aware Dynamic Voltage Scaling for Reliable Low Power Cache Memories

Daniele Rossi*, Vasileios Tenentes*, Saqib Khursheed†, Bashir M. Al-Hashimi*
*ECS, University of Southampton, UK. Email: {D.Rossi, V.Tenentes, bmah}@ecs.soton.ac.uk
†Electrical Engineering & Electronics, University of Liverpool, UK. Email: S.Khursheed@liverpool.ac.uk

Abstract—**We propose a novel dynamic voltage scaling (DVS) approach for reliable and energy efficient cache memories. First, we demonstrate that, as memories age, leakage power reduction techniques become more effective due to sub-threshold current reduction with aging. Then, we provide an analytical model and a design exploration framework to evaluate trade-offs between leakage power and reliability, and propose a BTI and leakage aware selection of the "drowsy" state retention voltage for DVS of cache memories. We propose three DVS policies, allowing us to achieve different power/reliability trade-offs. Through SPICE simulations, we show that a critical charge and a static noise margin increase up to 150% and 34.7%, respectively, is achieved compared to standard aging unaware drowsy technique, with a limited leakage power increase during the very early lifetime, and with leakage energy saving up to 37% in 10 years of operation. These improvements are attained at zero or negligible area cost.**

I. Introduction

Power has become a major concern for modern processor design due to thermal dissipation limitations of packaging and cooling [8]. As technology shrinks, leakage power is increasing dramatically, to the point where it can be nearly as large as dynamic power [8]. SRAMs are responsible for an important portion of the total chip leakage power consumption [13], because they occupy a large area of the chip. As an example, large L2/L3 cache memories in recent multicore processors occupy a large portion of the die, and they potentially represent a big source of leakage power, since they may remain unaccessed for long periods [10].

Power gating and dynamic voltage scaling (DVS) are two leakage power reduction techniques for memories [8], [16]. Although power gating is more effective in saving leakage power, it does not support data retention. Therefore, in power-gated cache memories, data are retrieved from upper level memories in the memory hierarchy, with performance penalties and the risk to undermine the energy savings of power gating.

On the other hand, DVS guarantees data retention, but is less effective for leakage power saving. Among DVS solutions [4], [6], [8], [9], the drowsy technique is proposed for on-chip caches [7], and is the focus of this paper. According to drowsy DVS, cache lines that are not being accessed are set into a low voltage mode (*drowsy mode*). During drowsy mode, the cache state is preserved, so there is no need to reload data from upper level memories. Therefore, the drowsy cache technique can allow up to 75% of energy reduction with no more than 1% of performance overhead [7], [12].

The low voltage of drowsy mode, denoted as *drowsy voltage* V_{dd}^{D}, degrades the reliability of the memory compared to active mode, and a memory cache line could stay in drowsy mode for a big portion of its lifetime [10]. Indeed, soft error susceptibility increases substantially due to critical charge Q_{crit} reduction when supply voltage is reduced [5]. Moreover, memory robustness to noise decreases due to static noise margin (SNM) reduction [11].

Both soft error susceptibility and SNM of low-power memories are further undermined by device aging. Bias temperature instability (BTI), whose main effect is to increase MOS transistor threshold voltage (V_{th}), is considered the primary parametric failure mechanism for nanometer CMOS technology [2], [17]. In [11], the negative effect of aging on memory reliability has been considered for the selection of the minimum voltage that guarantees high reliable data retention in low-power memories. However, this technique ignores the positive effect of BTI-induced aging on the sub-threshold current reduction, as shown in [15].

In this paper, to the best of our knowledge, we are the first to show that BTI-induced degradation can considerably benefit leakage power saving of drowsy cache memories, and we propose a BTI and leakage aware DVS approach for reliable low-power cache memories. In Sec. II, we review drowsy technique and BTI. In Sec. III, we first propose a DVS aware aging analytical model allowing us to properly account for the degradation of a drowsy memory and, based on that, we assess the BTI impact on a drowsy memory cell, considering a standard drowsy cache design. Through SPICE simulations, we show that leakage power may reduce by more than 35% during the first month of operation, by more than 48% during the first year, and up to 61% in 10 years of memory operation, considering a drowsy cache memory cell implemented in a 32nm, Metal Gate, High-K, Strained-Si CMOS technology [1]. Based on the proposed analytical model, in Sec. IV we develop a design exploration framework allowing us to evaluate several possible trade-offs between power consumption and reliability. Then, in Sec. V, we derive three drowsy voltage selection policies, each characterized by a different leakage power and reliability trade-off. Through SPICE simulations, we show this improves soft error resilience and SNM during drowsy mode, compared to a standard drowsy cache technique, exhibiting a Q_{crit} and SNM increase up to 150% and 34.7%, respectively. A very limited increase in leakage power consumption, compared to the value expected by a standard, BTI-unaware drowsy technique is exhibited during only the very early lifetime, while a leakage energy

saving up to 37% for 10 years of operation is achieved. These improvements are attained at zero or very limited area overhead (estimated under 3% for a 64 byte size cache memory line). Finally, in Sec. VI we draw some conclusions.

II. Background

Bias temperature instability causes a threshold voltage increase in MOS transistors, denoted by ΔV_{th}, when they are ON (stress phase) [3]. BTI-induced degradation is partially recovered when MOS transistors are polarized in their OFF state (recovery phase). Negative BTI (NBTI) is observed in pMOS transistors, and it usually dominates against the positive BTI (PBTI) observed in nMOS transistors [3]. The reaction-diffusion model in [3] allows designers to estimate ΔV_{th} as a function of technology parameters, operating conditions and time. Since ΔV_{th} does not depend on the frequency of input signals, but only on the total amount of the stress time, in [17] a simple analytical model has been proposed that allows designers to estimate long term threshold voltage shift. It is:

$$\Delta V_{th} = \chi K \sqrt{C_{ox}(V_{dd} - V_{th})} \alpha^n t^n \tag{1}$$

The parameter C_{ox} is the oxide capacitance, t is the operating time, and α is the fraction of the operating time during which a MOS transistor is under a stress condition. It is $0 \leq \alpha \leq 1$, where $\alpha = 0$ if the MOS transistor is always OFF (recovery phase), while $\alpha = 1$ if it is always ON (stress phase). The exponent $n = 1/6$ is a fitting parameter; the coefficient χ allows us to distinguish between PBTI and NBTI. Particularly, χ equals 0.5 for PBTI, and 1 for NBTI. The parameter K lumps technology specific and environmental parameters, and has been estimated to be $K \simeq 2.7V^{1/2}F^{-1/2}s^{-n}$ by fitting the model with the experimental results reported in [18].

Drowsy cache is a promising approach to reduce leakage power of cache cells, yet retaining their state, based on DVS [7]. When a cache line is not accessed, it is put into a low-power drowsy mode, thus reducing considerably the associated leakage power consumption ($P_{leak} = V_{dd} I_{leak}$). The high voltage level is restored before cache line content is accessed.

Leakage current I_{leak} has two main contributors [8]: sub-threshold current and gate current. Sub-threshold current contribution dominates, since gate current can be well controlled by the use of high-k dielectrics. Therefore, in a first order approximation [8], MOS transistor leakage current I_{leak} is:

$$I_{leak} \simeq \mu C_{ox} \left(\frac{kT}{q}\right)^2 \frac{W}{L} e^{\frac{-q(V_{gs} - V_{th})}{mkT}}. \tag{2}$$

If V_{dd} (V_{GS}) reduces, I_{leak} decreases as well, so does P_{leak}. In standard drowsy caches [7] the low V_{dd} value employed during the drowsy mode is determined in order to considerably reduce P_{leak}, yet being able to retain the memory state, without considering BTI-induced degradation. It is approximately equal to $1.5\times$ the value of the threshold voltage of memory cell transistors [7], a value that guarantees a good leakage power reduction, yet providing the design with adequate margins against noise and process variations [7]. Therefore, designers identify an expected leakage power consumption in drowsy mode as a target, which remains constant for the whole memory lifetime.

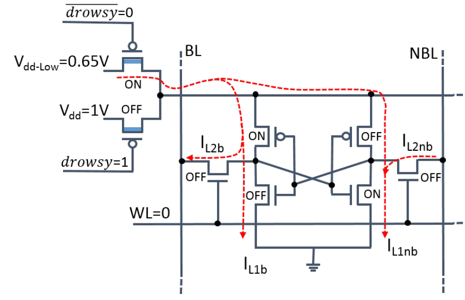

Fig. 1. Drowsy memory cell [7].

III. Analysis of BTI Impact on a Drowsy Memory Leakage Power and Reliability

In order to assess the impact of BTI on a drowsy memory, we considered the memory cell scheme shown in Fig. 1. It has been implemented in a 32nm Metal Gate, High-K Strained-Si CMOS technology [1], with a supply voltage (during active mode) $V_{dd} = 1V$. Particularly, the *high* V_{th} *low power* model (denoted by V_{th}^H) has been adopted to implement the pMOS power switches connected to the power supplies, as suggested in [7], while all other transistors have been designed using the *low* V_{th} *high performance* model (denoted by V_{th}^L). The value of the drowsy voltage is set to V_{dd}^D=0.65V, which is approximately equal to $1.5 \times V_{th}^L$ [7]. In Fig. 1 the leakage current paths are also highlighted (dashed arrows).

A. DVS Aware Aging Model for Drowsy Cache Memories

When a cache line switches to drowsy mode, its supply voltage is reduced, thus decreasing BTI degradation compared to active mode. Therefore, to properly estimate the BTI degradation of a memory cell, we modified the model in (1) to account for the different degradation induced during active mode and drowsy mode. Let us define as *access ratio* the ratio between the total operating time and the time during which the considered cache line is operating in active mode, and denote it by γ. In turn, the ratio of the operating time during which the memory is operating in drowsy mode is $(1 - \gamma)$. Note that the power switch connected to the drowsy V_{dd} and the transistors composing a memory cell experience a different stress time. Given α the stress time ratio (Sec. II), the new aging model formulation for the V_{th}^L transistors composing a drowsy memory cell is:

$$\Delta V_{th}^L = \chi K \Big\{ \gamma \sqrt{C_{ox}(V_{dd} - V_{th}^L)} + (1-\gamma)\sqrt{C_{ox}(V_{dd}^D - V_{th}^L)} \Big\} \alpha^n t^n. \tag{3}$$

The V_{th}^H pMOS power switch connected to the drowsy V_{dd} is exposed to a stress time with a ratio $\alpha = (1-\gamma)$. Therefore, the aging model for this transistor is:

$$\Delta V_{th}^H = K \sqrt{C_{ox}(V_{dd}^D - V_{th}^H)} (1-\gamma)^n t^n. \tag{4}$$

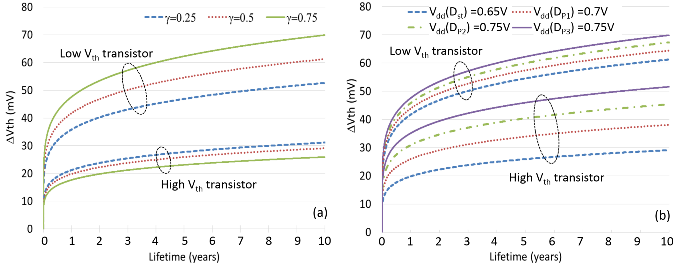

Fig. 2. Threshold voltage degradation profile over time for both low V_{th} and high V_{th} transistors, as a function of: (a) access ratio γ (V_{dd}^D = 0.65V); (b) V_{dd}^D ($\gamma = 0.5$).

In Fig. 2, we depict the trend over time of the threshold voltage degradation of the memory cell transistors, as given by (3), and of the power switch connected to the drowsy V_{dd}, as given by (4). The value of the stress ratio α has been set equal to 0.5, and values 0.25, 0.5 and 0.75 have been considered for the access ratio γ, as highlighted in Fig. 2(a). Note that cell transistors (*Low* V_{th}) experience a higher degradation compared to power switch (*High* V_{th}) connected to the drowsy V_{dd}. Moreover, the degradation of memory cell transistors increases with γ, since larger γ values represent longer time periods during which the memory operates in active mode (powered with $V_{dd} = 1V$) and is subjected to a larger stress. On the other hand, the degradation of the power switch connected to V_{dd}^D decreases with γ, since the stress ratio for this transistor is given by $(1 - \gamma)$. In Fig. 2(b), the trend over time of ΔV_{th} for different values of V_{dd}^D (0.65V, 0.7V, 0.75V and 0.8V) is shown. As expected, the degradation increases with voltage, and this increase is more evident for the high V_{th} power switch than the low V_{th} cell transistors.

B. BTI-Induced Degradation of Soft Error Susceptibility and SNM During Drowsy Mode

DVS increases memory soft error susceptibility and reduces SNM [5]. As a result, drowsy memories are much more susceptible to reliability threats when operated in drowsy mode than in active mode. Therefore, we assess the BTI-induced degradation of soft error susceptibility and SNM of a cache memory, when it operates in drowsy mode.

Soft error susceptibility is evaluated by considering the critical charge Q_{crit}, which is defined as the minimum amount of charge collected by a node that is able to flip the affected memory cell. In drowsy mode, Q_{crit} reduces by more than 87% compared to active mode (from 10.4fC to 1.3fC at $t0$). Moreover, Q_{crit} is further degraded by BTI. To evaluate Q_{crit} profile over time, we estimate ΔV_{th} by (3) for cell transistors and (4) for power switches. Similarly to [14], [18], the estimated ΔV_{th} values for each considered lifetime have been utilized to customize the SPICE device model, so that each transistor is simulated with the proper BTI degradation. In Fig. 3, the Q_{crit} values for a memory lifetime up to 10 years are shown for different values of access ratio γ. The relative Q_{crit} reductions with respect to $t0$ value are also shown. Note that the Q_{crit} decreases by more than 26% over 10 years, reaching 20% reduction after only 1 year. Moreover, Q_{crit}

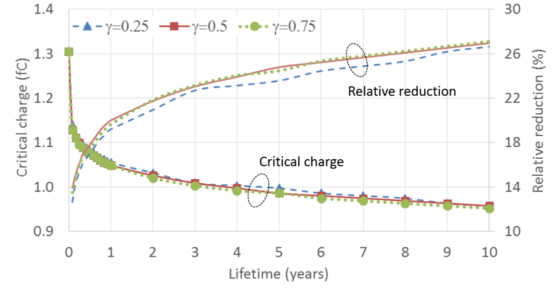

Fig. 3. Critical charge profile over time for the considered values of access ratio γ and V_{dd}^D)=0.65V, and relative reduction with respect to Q_{crit} at $t0$: $Q_{crit}(t0) - Q_{crit}(t)]/Q_{crit}(t0)]$.

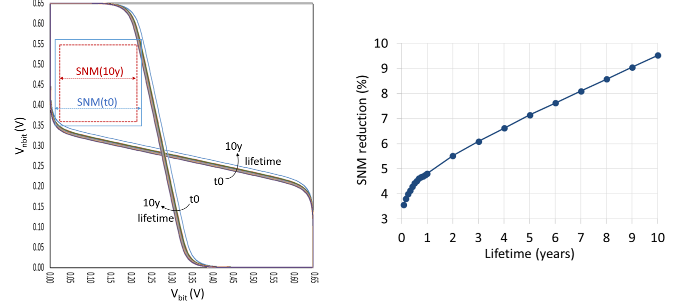

Fig. 4. SNM for trend over time $V_{dd}(D_{st})$=0.65 and access ratio $\gamma = 0.5$: (a) butterfly plot; (b) SNM reduction over time with respect SNM at $t0$: $[SNM(t0) - SNM(t)]/SNM(t0)]$

slightly depends on access ratio γ, despite the fact that the threshold voltage degradation shows an evident dependence on it. This can be attributed to the opposite dependence of degradation of cell transistors and power switch on γ (Fig. 2(a)). The Q_{crit} reduction impact is greater compared to that exhibited by standard SRAM cell operating with $V_{dd} = 1V$. For this latter we found a 11.4% Q_{crit} reduction over 10 years, in line with the values reported also in [14]. This difference (26% to 11.4%) can be attributed to the presence of the power switch, whose BTI degradation exacerbates the Q_{crit} reduction.

As for SNM, we found that, in drowsy mode, it is reduced to less than 56% of that of active mode (from 376mV to 210mV at $t0$). Moreover, similarly to the case of Q_{crit}, BTI-induced degradation further decreases SNM over time. SNM profile has been obtained graphically by means of the butterfly plot, and the SPICE simulation results are depicted in Fig. 4. The SNM reduces by 9.5% over ten years of operation, thus exhibiting a degradation over time considerably lower than Q_{crit}. No appreciable impact of access ratio γ was found.

C. BTI Impact on Leakage Power during Drowsy Mode

For the considered case study, when a memory cell switches from active mode to drowsy mode, leakage power drops to $227pW$, with a reduction exceeding 94% with respect to a standard memory design with no DVS. This value represents the leakage power expected to be consumed by a standard drowsy technique not accounting for BTI. We will refer to this value as *expected leakage power at t0*, and we will denote it as EP_{leak0}. Instead, we expect that leakage power considerably decreases as memory ages [15]. This is confirmed by the

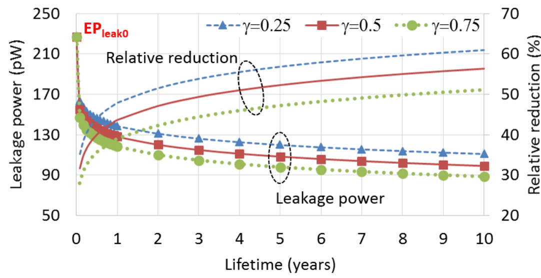

Fig. 5. Leakage power trend over time for the considered values of γ and $V_{dd}(D_{st})$=0.65V, and relative variation with respect to $t0$ values: $[P_{leak}(t0) - P_{leak}(t)]/P_{leak}(t0)]$.

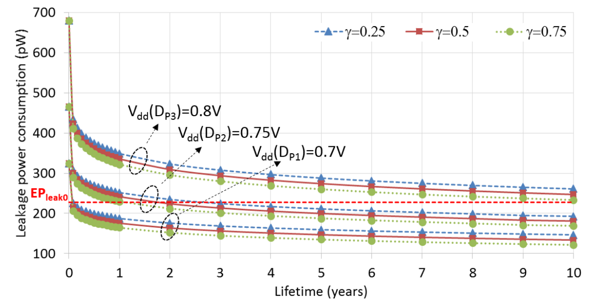

Fig. 6. Leakage power profile for a cache memory implementing drowsy modes D_{P1}, D_{P2} and D_{P3}, for the considered access ratio γ values.

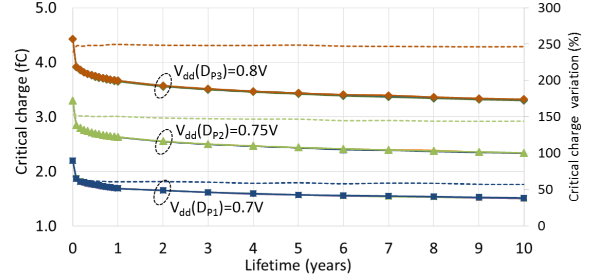

Fig. 7. Critical charge profile for a cache memory implementing drowsy modes D_{P1}, D_{P2} and D_{P3}, for the considered values of access ratio γ (solid lines), and variations over the standard drowsy memory D_{st} (dashed lines): $[Q_{crit}(D_{Pi}, t) - Q_{crit}(D_{st}, t)]/Q_{crit}(D_{st}, t)]$, for $i = (1, 2, 3)$.

simulation results shown in Fig. 5 for the considered values of access ratio γ (0.25, 0.5 and 0.75). The relative reduction over time is also shown. After only 1 month of operations, leakage power reduction ranges from 28% ($\gamma = 0.25$) to 35% ($\gamma = 0.75$); after 10 years, leakage power reduction reaches 51% for $\gamma = 0.25$, and 61% for $\gamma = 0.75$. We observe that, similarly to Q_{crit} and SNM, leakage power decreases after 1 month of operation by more than 50% of the variation exhibited after 10 years of operation. On the other hand, leakage power variation depends noticeably on access ratio γ. In particular, the leakage power variations for $\gamma = 0.25$ (lowest degradation, as shown in Fig. 2(a)) and $\gamma = 0.75$ (highest degradation) differ by 10%. This is attributed to the higher sensitivity of leakage power to V_{th} degradation compared to Q_{crit} and SNM. These two quantities are proportional to the driving strength (active current) of memory cell transistors, which depends almost linearly on the overdrive voltage $V_{gs} - V_{th}$. Instead, the sub-threshold leakage current, which is the dominant contributor to leakage power, varies exponentially with $V_{gs} - V_{th}$, as reported in (2). Finally, SPICE simulation results confirm that leakage power decreases over time to a value considerably lower than EP_{leak0} estimated by a standard, BTI-unaware drowsy technique, clearly showing the positive effect of aging on leakage power.

IV. Proposed Framework for Power & Reliability Aware DVS Design Exploration

The beneficial impact of aging on leakage power, which reduces over time well below the expected value EP_{leak0} has been ignored so far by DVS techniques. We propose to trade-off some of this leakage power over-reduction in order to counteract the detrimental effect of BTI aging on soft error susceptibility and SNM, thus improving memory reliability. This can be achieved by selecting a higher drowsy voltage to be applied to cache lines not being accessed. Of course, different drowsy voltage values enable to achieve different trade-offs between leakage power consumption and reliability. In this section, we develop a design exploration framework allowing designers to evaluate leakage power and reliability trade-offs. In this regard, we analyze the trend over time of P_{leak}, Q_{crit} and SNM considering three different drowsy modes, denoted by D_{P1}, D_{P2} and D_{P3} characterized by the following drowsy supply voltages, all higher than the value of the standard drowsy technique ($V_{dd}(D_{st}) = 0.65V$): $V_{dd}(D_{P1}) = 0.7V$, $V_{dd}(D_{P2}) = 0.75V$ and $V_{dd}(D_{P3}) = 0.8V$.

In Fig. 6, we show the P_{leak} profile for a cache memory implementing drowsy modes D_{P1}, D_{P2} and D_{P3}, and for the three considered values of access ratio γ. Similarly to the results depicted in Fig. 5, P_{leak} decreases rapidly for all values of γ. As expected, P_{leak} values at $t0$ are higher than EP_{leak0} (dashed red line in Fig. 6). However, in the case of drowsy mode D_{P1}, P_{leak} drops below EP_{leak0} after less than a month of operation for all values of γ. For the drowsy mode D_{P2}, instead, EP_{leak0} is reached after 1.2 years for $\gamma = 0.75$, 2.7 years for $\gamma = 0.25$, 1.8 years for $\gamma = 0.5$. Finally, for the drowsy mode D_{P3}, EP_{leak0} is approximated only for $\gamma = 0.75$ after 10 years of operation.

Fig. 7 shows the Q_{crit} profile over time for the considered scenarios and access ratio γ, together with the respective variations over the standard drowsy technique D_{st} (dashed lines). Q_{crit} profiles for different γ are completely overlapped. As expected, the Q_{crit} increases noticeably with the increase of drowsy V_{dd}. The Q_{crit} improvement over D_{st} ranges from 50% for the D_{P1} scenario to approximately 250% for the D_{P3} scenario. Moreover, we can observe that the Q_{crit} improvement slightly varies over time.

In Table I, we report the SNM values for the considered scenarios for several lifetime values, together with the respective variation over the SNM provided by the standard drowsy memory D_{st}. As can be seen, the provided SNM improvement over the standard approach ranges from 11.1% for the D_{P1} scenario to 34.7% for the D_{P3} scenario.

So far, we have addressed the analysis of the impact of

TABLE I
SNM VALUES AND VARIATION OVER A STANDARD DROWSY TECHNIQUE
$(\Delta = [(SNM(D_{Pi}, t) - SNM(D_{st}, t)]/SNM(D_{st}, t), i = 1, 2, 3)$

Lifetime	$V_{dd}(D_{P1}) = 0.7V$		$V_{dd}(D_{P2}) = 0.75V$		$V_{dd}(D_{P3}) = 0.8V$	
	SNM (mV)	Δ%	SNM (mV)	Δ%	SNM (mV)	Δ%
t0	235	11.9	258	22.9	282	34.3
1m	225	11.4	245	21.3	270	33.7
1y	222	11.6	242	21.6	267	34.2
10y	211	11.1	233	22.6	256	34.7

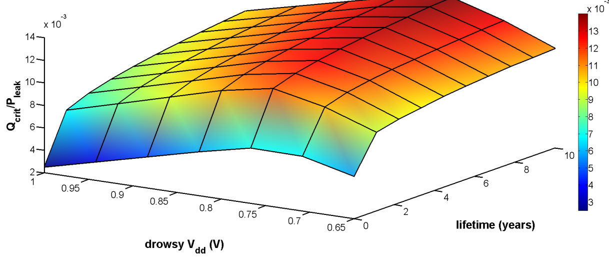

Fig. 8. Reliability power efficiency metric profile over time for the considered drowsy voltage modes D_{P1}, D_{P2} and D_{P3} and access ratio $\gamma = 0.75$.

the considered drowsy modes on either leakage power or reliability features (Q_{crit} and SNM) separately. We now define a new metric allowing us to jointly evaluate reliability and leakage power consumption. Particularly, we focus on Q_{crit} as a reliability aspect, which has been found to be much more dependent on the adopted drowsy V_{dd} and to degrade much more than SNM with aging. The new metric, defined as Q_{crit}/P_{leak}, represents the critical charge offered by a solution per unit of leakage power consumed. It is therefore an evaluation of the power efficiency in providing resilience against soft errors during drowsy mode. It is depicted in Fig. 8 as a function of drowsy voltage and lifetime. As we can see, the Q_{crit}/P_{leak} metric increases over time for all considered cases. Indeed, as discussed in Sec. III, P_{leak} decreases faster with lifetime compared to Q_{crit}. Moreover, the depicted function exhibits a maximum for $V_{dd}(D_{P2}) = 0.75V$ for all lifetime values. This can be explained by considering that P_{leak} increases exponentially with V_{dd}, while Q_{crit} is almost linear with it. If for small value of the drowsy V_{dd} the Q_{crit}/P_{leak} metric is benefited by an increase of power supply, larger drowsy V_{dd} values turn-out to be a power inefficient approach for soft error resilience increase.

V. Proposed DVS Policies for Reliable Low Power Cache Memories and Validation Results

From the simulation results obtained with the proposed design exploration framework, we derive and evaluate three different drowsy V_{dd} selection policies, leading to three different power and reliability trade-offs. They are: 1) static selection of a drowsy power supply suitably higher than in the standard approach (equal to 0.65V) in order to increase memory reliability yet meeting leakage power/energy constraints, referred to as *Upgraded Drowsy* and denoted by *UD*; 2) dynamic (adaptive) selection of drowsy V_{dd} over time, in order to further increase reliability compared to *UD*, yet meeting leakage power/energy constraints, referred to as *Upgraded Adaptive Drowsy*, and

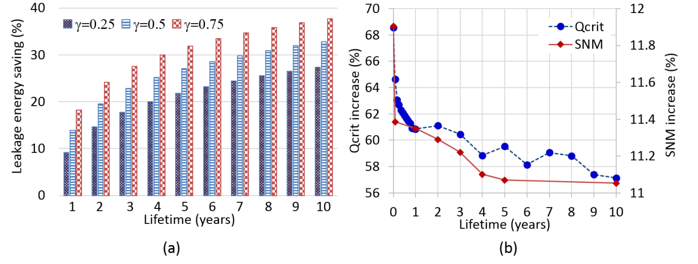

Fig. 9. *UD*: variation over time of (a) leakage energy and (b) Q_{crit} and SNM, over the standard drowsy technique.

denoted by *UAD*; 3) selection of a drowsy V_{dd} in order to maximize the Q_{crit}/P_{leak} metric, as defined in Sec. IV, referred to as *Reliable power Efficient Drowsy*, and denoted by *RED*. The proposed drowsy V_{dd} selection policies have been validated through SPICE simulations by evaluating the leakage energy saving with respect to the value expected for a standard, BTI-unaware drowsy technique. Moreover, Q_{crit} and SNM variation over the standard drowsy technique have been also considered as metrics for comparison, and evaluated as $[A(D_{pi}, t) - A(D_{st}, t)]/A(D_{st}, t)]$, with $A = (Q_{crit}, SNM)$ and $i = 1, 2, 3$.

In the *UD* policy, a drowsy power supply $V_{dd}(D_{P1}) = 0.7V$ is selected. Fig. 9 depicts the obtained simulation results. As can be seen, in 10 years of operation the energy saving (Fig. 9(a)) ranges from 26% for $\gamma = 0.25$ to 38% for $\gamma = 0.75$. As for the Q_{crit} improvement over time (Fig. 9(b)), it ranges from 68% at $t0$ to 57% at 10 years, while SNM increase is in the interval 11%-12% for all lifetime values. It is worth noticing that the *UD* does not introduce any hardware overhead over the standard drowsy technique.

If the *UAD* policy is adopted, the memory switches from drowsy mode D_{P1} ($V_{dd}(D_{P1}) = 0.7V$) to drowsy mode D_{P2} ($V_{dd}(D_{P2}) = 0.75V$) during its lifetime, in order to further improve reliability compared to the *UD*, yet meeting the leakage/power energy constraint. The selection over time of the proper drowsy V_{dd} can be driven by a control signal provided by an already present aging monitor, or generated at system level. Considering the simulation results shown in Fig. 6 (Sec. IV), the switching time from D_{P1} to D_{P2} has been set at the fourth year, which allows us to meet the expected leakage power constraint for all considered values of access ratio γ. Fig. 10(a) shows the leakage energy saving over a standard drowsy technique. When the drowsy mode switches from D_{P1} to D_{P2}, the energy saving over the expected value reduces, and then increases again up to 20% (for $\gamma = 0.75$) after 10 years of operation. Meanwhile, the Q_{crit} (SNM) improvement over time (Fig. 10(b)) increases from around 60% (11%) during the first 3 years, to slightly less than 150% (25%) for the rest of lifetime. Compared to the *UD*, a higher soft error resilience over time is achieved at the cost of less energy saving over the standard drowsy technique. The described reliability improvement comes together with a small hardware cost, since this approach requires the on-chip generation of 2 different drowsy V_{dd}, one additional power switch and an *ad-hoc* control logic per cache line. The

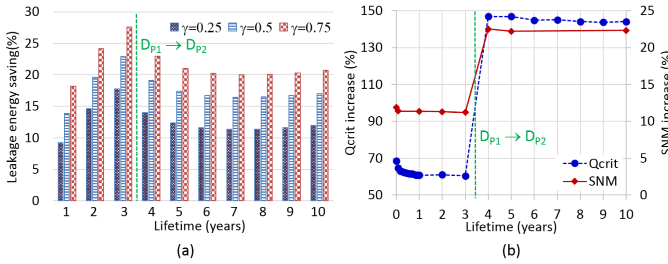

Fig. 10. *UAD*: variation over time of (a) leakage energy and (b) Q_{crit} and SNM, over the standard drowsy technique.

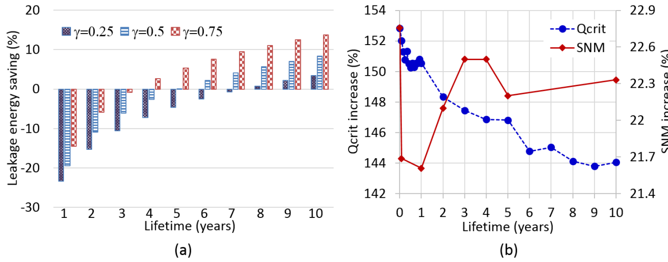

Fig. 11. *RED*: variation over time of (a) leakage energy and (b) Q_{crit} and SNM, over the standard drowsy technique.

detailed design and evaluation of this additional circuit is out of the scope of this paper. Roughly estimating its hardware overhead over the standard drowsy memory technique in terms of transistor count, it is lower than 3% for a cache memory with a 64 byte line size.

In the *RED* policy, the drowsy voltage $V_{dd}(D_{P2}) = 0.75V$ is selected in order to maximize the the Q_{crit}/P_{leak} metric, thus the power efficiency in providing drowsy memory with soft error resilience. From the simulation results in Fig. 11, we can see that the Q_{crit} (SNM) increase with respect to the standard drowsy technique is in the range 144%-153% (21.6%-22.7%) over the whole lifetime. This noticeable reliability improvement is achieved at the cost of an increase in leakage energy consumption for the first 4 years of operation over the standard drowsy technique, but with no hardware overhead.

VI. Conclusions

We have shown that BTI-induced degradation can considerably benefit leakage power saving of drowsy cache memories. We developed an analytical model and a design exploration framework allowing us to evaluate several trade-offs between power consumption and reliability, and proposed a BTI and leakage aware selection of the drowsy voltage for DVS of cache memories. Finally, we proposed three DVS policies, allowing us to achieve different power/reliability trade-offs. Through SPICE simulations, we showed that, compared to standard aging unaware drowsy technique, a critical charge improvement up to 150% and a static noise margin increase up to 34.7% is enabled, with a limited increase in leakage power during only the very early lifetime, and with leakage energy saving up to 37% in 10 years of operation. These improvements are attained at no or very limited area overhead, estimated under 3% for a 64 byte size cache memory line.

Acknowledgments

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.

References

[1] "Predictive Technology Model (PTM)," http://www.ptm.asu.edu.
[2] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. C. Paul, W. Wang, B. Yang, Y. Cao, and S. Mitra, "Optimized circuit failure prediction for aging: Practicality and promise," in *Proc. of IEEE International Test Conf. (ITC)*, 2008, pp. 1–10.
[3] M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, "A comprehensive model for pmos nbti degradation: Recent progress," *Microelectronics Reliability*, vol. 47, no. 6, pp. 853–862, 2007.
[4] A. Bardine, M. Comparetti, P. Foglia, and C. A. Prete, "Evaluation of leakage reduction alternatives for deep submicron dynamic nonuniform cache architecture caches," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 22, no. 1, pp. 185–190, 2014.
[5] V. Chandra and R. Aitken, "Impact of technology and voltage scaling on the soft error susceptibility in nanoscale cmos," in *Defect and Fault Tolerance of VLSI Systems, 2008. DFTVS'08. IEEE International Symposium on*. IEEE, 2008, pp. 114–122.
[6] H. David, C. Fallin, E. Gorbatov, U. R. Hanebutte, and O. Mutlu, "Memory power management via dynamic voltage/frequency scaling," in *Proceedings of the 8th ACM international conference on Autonomic computing*. ACM, 2011, pp. 31–40.
[7] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power," in *Computer Architecture, 2002. Proceedings. 29th Annual International Symposium on*. IEEE, 2002, pp. 148–157.
[8] D. Flynn, R. Aitken, A. Gibbons, and K. Shi, *Low Power Methodology Manual: For System-on-Chip Design*. NY, USA: Springer-Verlag, 2007.
[9] M. J. Geiger, S. A. McKee, and G. S. Tyson, "Drowsy region-based caches: minimizing both dynamic and static power dissipation," in *Proceedings of the 2nd conference on Computing frontiers*. ACM, 2005, pp. 378–384.
[10] N. S. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 167–184, 2004.
[11] T. T.-H. Kim and Z. H. Kong, "Impact analysis of nbti/pbti on sram v min and design techniques for improved sram v min," *JSTS: Journal of Semiconductor Technology and Science*, vol. 13, no. 2, pp. 87–97, 2013.
[12] M. Kulkarni, K. Sheth, and V. D. Agrawal, "Architectural power management for high leakage technologies," in *System Theory (SSST), 2011 IEEE 43rd Southeastern Symposium on*. IEEE, 2011, pp. 67–72.
[13] A. Nourivand, A. J. Al-Khalili, and Y. Savaria, "Postsilicon tuning of standby supply voltage in srams to reduce yield losses due to parametric data-retention failures," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 20, no. 1, pp. 29–41, 2012.
[14] D. Rossi, M. Omaña, C. Metra, and A. Paccagnella, "Impact of aging phenomena on soft error susceptibility," in *Proc. of IEEE International Symp. on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 2011, pp. 18–24.
[15] 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 *ETS'15, to appear*, http://eprints.soton.ac.uk/374987/1/ets15-84.pdf.
[16] J. Wang and B. H. Calhoun, "Minimum supply voltage and yield estimation for large srams under parametric variations," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 11, pp. 2120–2125, 2011.
[17] W. Wang, Z. Wei, S. Yang, and Y. Cao, "An efficient method to identify critical gates under circuit aging," in *Proc. of IEEE/ACM International Conf. on Computer-Aided Design (ICCAD)*, 2007, pp. 735–740.
[18] 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.