INTEGRATION the VLSI journal xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Lifetime improvement by exploiting aggressive voltage scaling during runtime of error-resilient applications

Farzaneh Nakhaee[a], Mehdi Kamal[a,*], Ali Afzali-Kusha[a,b], Massoud Pedram[c], Sied Mehdi Fakhraie[a], Hamed Dorosti[a]

[a] School of Electrical and Computer Engineering, University of Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran
[c] Department of Electrical Engineering, University of Southern California, USA

ARTICLE INFO

Keywords:
Dynamic voltage scaling
Aging mechanisms
Lifetime improvement
Quality degradation

ABSTRACT

In this paper, we present an accuracy-aware operating voltage management unit to improve the lifetime of processors by considering the error-resilient nature of some applications. This unit is placed in the power management unit of the processor and its operation is based on the aggressive operating voltage reduction during the runtime of error resilient applications. This unit determines the operating voltage of the processor based on the type of the running application, the predefined minimum acceptable quality, and the operating voltage level specified by the dynamic voltage frequency scaling controller of the processor power management unit. The determined operating voltage by this unit results in lifetime improvement and power reduction at the cost of timing violations that are tolerable by error-resilient applications. In addition, the proposed unit dynamically adjusts the minimum acceptable operating voltage based on the impact of aging mechanisms. The aging mechanisms considered in this work include Negative Bias Temperature Instability, Hot Carrier Injection, Time Dependent Dielectric Breakdown, Thermal Cycling, Electro-Migration, and Stress Migration. The efficacy of the proposed operating voltage management is investigated by applying it to some exact and error-resilient applications. The results show that the proposed unit lead to, on average, 38.82% lifetime improvement as well as 41.8% power consumption reduction.

1. Introduction

Aggressive technology scaling has improved the performance of integrated circuits. However, adverse effects of the feature size reduction on MOS device operation are inevitable [1,2]. In particular, physical scaling increases the impact of process and temporal variations. One type of temporal variations, known as aging which is caused by different mechanisms, impacts the device functionality over time, reducing the circuit lifetime [3]. A number of methods at different abstraction levels have been suggested to enhance the system lifetime. Lower level methods, which are applied at the design time, change transistor parameters to reduce the aging rate (*e.g.*, [4,5]). Higher level methods, however, protect the system functionality from run-time failures due to temporal variations (*e.g.*, [2,6]). As higher level methods may be applied to large integrated circuits [7] and dynamically adjust the system parameters, they could be more efficient in terms of lifetime improvement [7]. These methods use the supply voltage and working frequency as the most important knobs to adjust the circuit performance to achieve higher lifetimes [2]. Therefore, suitable voltage and frequency assignment according to performance requirement is necessary to improve the circuit lifetime.

Different aging mechanisms affect the device lifetime. Negative Bias Temperature Instability (NBTI), Hot Carrier Injection (HCI), and Time Dependent Dielectric Breakdown (TDDB) are the most important mechanisms impacting the transistor parameters during its lifetime [8]. The NBTI mechanism is caused by the generation of interface traps at Si/SiO_2 interface [9]. The hot electrons/holes at the drain side can also be injected and trapped in the gate oxide (HCI effect) [10]. Trap changes that exist at the Si-oxide interface as well as inside the oxide will result in some threshold voltage change. If the change increases the absolute value of the threshold voltage, the circuit performance will be degraded by increasing the critical path delay. This may eventually cause a clock period violation adversely affecting the proper functions of a circuit. The TDDB degrades the quality of the gate dielectric over time creating a resistive path from the channel to the gate through the oxide layer and finally causing an oxide breakdown [10]. In addition,

* Corresponding author.
E-mail addresses: f.nakhaee@ut.ac.ir (F. Nakhaee), mehdikamal@ut.ac.ir (M. Kamal), afzali@ut.ac.ir (A. Afzali-Kusha), pedram@usc.edu (M. Pedram), fakhraie@ut.ac.ir (S.M. Fakhraie), hdorosti@ut.ac.ir (H. Dorosti).

https://doi.org/10.1016/j.vlsi.2017.10.013
Received 10 October 2017; Received in revised form 23 October 2017; Accepted 29 October 2017

power up, power down, and workload changes can cause cracks in the ball-bond and pads due to different thermal expansions of these materials. This effect is called Thermal Cycling (TC) [10]. Increasing the current density may cause the metal ions to be transported in the direction of the current flow eventually leading to short and open circuit failures due to the Electro-Migration (EM) failure [10]. Finally, Stress Migration (SM) occurs when a mechanical stress results in a plastic deformation due to the flow of metal ions to reach stress-relief [10]. All these aging mechanisms affect the device lifetime and should be considered by the designers.

Aging mechanisms generally are strong functions of temperature [11,12]. Therefore, in digital circuits, the chip temperature should be controlled. To reduce the temperature, the power consumption of the circuits should be reduced. By reducing the operating voltage, the power consumption is considerably reduced [13]. The voltage reduction also lowers the strength of the electric fields in the transistor, reducing the aging rate due to the fact that some of the aging mechanisms are electric field dependent. Since lowering the operating voltage increases the delay of the logic paths in the circuit, the operating frequency should be lowered as well to prevent any timing violation and malfunctioning of the circuit.

There is a class of applications that are error-resilient and have inherent error-tolerant characteristics. In these applications, some levels of errors including those induced by timing violations (which can in turn cause output quality degradation) may be tolerated. For this class of applications, reducing the supply voltage below the minimum voltage (which no timing violation occurs) would be possible. This type of design belongs to the category of *aggressive deployment* or *better than worst case* designs [14]. Approximate computing has been widely considered as a technique for reducing the power consumption and/or increasing the speed of a circuit. In [13], by defining an approximate instruction set architecture (ISA) and using lower supply voltage level for these instructions, the power and energy consumptions of the processor during the running of the error resilient applications were reduced. Because the circuit lifetime is one of the most important criteria in advanced technologies, improving the circuit lifetime using approximate computing may be employed as an effective solution. Although there are many works focusing on reducing the power/energy consumption by employing the approximate computing (*e.g.*, [13,15,16]), to the best of our knowledge, lifetime improvement by employing the approximate computing approach has not been considered in the prior works.

In this paper, an accuracy-aware Operating Voltage Management (OVM) technique which lowers the circuit aging rate without sacrificing its performance is presented. To increase the lifetime of the system, the proposed unit reduces the voltage determined by the dynamic voltage and frequency scaling (DVFS) module of the processor power management unit to a smaller value if and when the input application can tolerate timing violations. The tolerable error is specified based on the acceptable output quality of the application. The DVFS module selects the proper voltage and its corresponding frequency to reduce the circuit power consumption, whereas the proposed OVM unit reduces only the chosen voltage (without changing the frequency) to reach a higher lifetime without performance loss. The reduced voltage levels should be determined dynamically for each approximate application based on its minimum acceptable output quality. It is worth mentioning that due to the aging mechanisms, the minimum acceptable voltage to satisfy the minimum satisfactory output quality is changed during the lifetime. Hence, the proposed OVM unit has the ability to dynamically adjust the minimum operating voltage during the runtime to guarantee the minimum acceptable output quality.

Here, without loss of generality, image processing applications are considered as error-resilient applications. Image processing algorithms are widely used in many areas including medical imaging and pattern recognition. Due to their heavy workloads and hence possible high operating temperatures, aging is an important concern for these applications. The aging mechanisms considered in this work include NBTI, HCI, TDDB, TC, EM, and SM.

The contributions of this paper may be summarized as follows:

1. Employing the error-resiliency of applications to increase the lifetime of the system without losing performance.
2. Proposing an accuracy-aware OVM unit to reach higher lifetime and lower power consumption.
3. Proposing a heuristic process to choose the minimum acceptable operating voltage during the application runtime based on the satisfactory output quality.
4. Considering the aging mechanism impacts during the system lifetime and minimum satisfactory output quality constraint for adjusting the operating voltage.

The rest of the paper is organized as follows. In Section 2, a brief review of some works dealing with approximate computing based on voltage over-scaling is provided. In the next section, the proposed accuracy-aware OVM unit is explained while the aging models used in this work are discussed in Section 4. The simulation framework used to obtain the results is described in Section 5 while the results are discussed in Section 6. Finally, the paper is concluded in Section 7.

2. Related works

In this section, first, some of the works which used voltage over-scaling (leading to some computation approximation) to reduce the power/energy consumption are reviewed. Next, the details of some of the works focused on reducing the impacts of the aging phenomena are discussed.

2.1. Voltage over-scaling technique

In [16], different approximate computing methodologies for energy efficient designs have been discussed. The methods include both circuit-level and algorithm-level techniques. The circuit-level techniques were based on, *e.g.*, approximate adder and multiplier circuits. In these designs, by reducing the number of transistors, the power consumptions of the designs were reduced. At the algorithm level, reducing the number of iterations in loops was considered as a method of reducing the energy consumption. In addition, bit width reduction by omitting the bits with small impact on the calculation accuracy (*e.g.*, LSBs) was invoked. Finally, voltage over-scaling was suggested as a technique for reducing the energy consumption at the cost of some timing errors. In [17], a low-energy Digital Signal Processing (DSP) which operates at a voltage level smaller than the critical voltage (minimum operating voltage without timing violation) was proposed. The authors also suggested an error-control approach to mitigate the output errors.

In [18], to reduce the energy consumption of implementing meta-functions in error-resilient applications, a voltage over-scaled algorithm was proposed. In the proposed approach, to reduce the timing violation at over-scaled voltages, the meta-functions were implemented using multiple isolated paths. The possible error accumulation problem, occurring during the voltage over-scaling, was overcome by introducing a correction cycle. To reduce the energy consumption, the authors of [19] suggested a method for statistically recognizing significant and insignificant computations where efficient implementations were invoked for each type. A Delay Predictor Block was employed to estimate the inputs which could activate the critical path. For these inputs, an extra cycle was generated such that the computation could be completed when the voltage was over-scaled. In the over-scaling technique discussed in [20], the timing violations do not occur for the crucial computations. This was performed by not applying the voltage over-scaling to the functions which should be performed without any error. This selective voltage over-scaling prevents con-