Integration, the VLSI Journal xxx (2018) 1–12

Contents lists available at ScienceDirect

Integration, the VLSI Journal

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

Approximate Error Detection-Correction for efficient Adaptive Voltage Over-Scaling

Roberto G. Rizzo [a,*], Andrea Calimera [a], Jun Zhou [b]

[a] Politecnico di Torino, Torino, Italy
[b] University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Keywords:
Error Detection-Correction
Energy optimization
Error-resilient applications
Adaptive Voltage Over-Scaling

ABSTRACT

This paper introduces *Approximate Error Detection-Correction* (AED-C), an error management scheme suited to adaptive power management on error resilient applications. Inspired by the working principle of Approximate Computing, AED-C implements energy-accuracy scaling using the error detection coverage as a knob: a low error coverage accelerates supply voltage scaling thus to achieve larger energy savings at the cost of quality-of-result (QoR); a high error coverage lessens the voltage scaling leading to high QoR at the cost of weaker energy savings. The AED-C mechanism is built upon *elastic timing monitors*, *Razor* flip-flops augmented with a tunable detection window and hardened with the aid of a dynamic short-path padding technique. Simulations over a representative set of circuits provide a comparative analysis with the state-of-art. The collected results show AED-C substantially reduces the average energy-per-operation (up to 44.7% savings w.r.t. Razor-driven Adaptive Voltage Over-Scaling) and the area overhead (3.3% vs. 62.0%), still guaranteeing reasonable QoR. When applied to a real-life application, i.e., Forward Discrete Cosine Transform Unit (FDCT) integrated into a JPEG compressor, AED-C shows 51.9% energy savings (w.r.t. a baseline FDCT implementation) and a PSNR of 48.45 dB (w.r.t. baseline JPEG images).

1. Introduction

1.1. Context

Adaptive power-management strategies represent a viable option to match the stringent energy constraints of portable/wearable applications [1]. They give circuits the ability to understand, at run-time, depending on the context, i.e. the flow of data, if/when there's margin to further reduce the energy consumptions below the theoretic minimum.

Adaptive Voltage Over-Scaling (AVOS) [2] represents a practical example of such strategies. It is based on the same working principle of the Dynamic Voltage Frequency Scaling (DVFS), yet with substantial differences. While DVFS brings the circuit close to the minimum energy by choosing the voltage-frequency pair on the basis of a static timing analysis [3], AVOS allows circuits to operate below the minimum energy exploiting information collected at run-time. Indeed, AVOS reduces the supply voltage depending on the actual workload and the longest *sensitized* timing path. This allows Vdd to approach values below the "safety" threshold defined at design-time and reach larger energy savings.

The use of AVOS implies the availability of on-chip monitoring architectures that provide the power manager with a real-time feedback on the health of the circuit. Among the existing schemes, those that use timing-errors, e.g. the *Razor* strategy [4–8], have proven quite effective. They leverage in-situ timing sensor to check whether the current workload incurred timing violations. The error detection is implemented through special Flip-Flops (FFs), called Razor-FFs, which sample the combinational output of a logic-cone at two different instants of time: first, at the rise edge of the clock, second, after a predefined timing window, the so-called *Detection Window* (DW). A parity check on the two time-skewed samples returns an error flag: a match implies a correct logic computation, and hence, the availability of some timing slack that can be consumed through Vdd scaling; a mismatch implies a set-up time violation, and hence, a timing error that can be eventually recovered using some correction mechanism. Since error correction is a costly procedure, a too fast voltage lowering could lead to a high error rate; this may induce latency/power penalties that overwhelm the savings

* Corresponding author.
E-mail addresses: robertogiorgio.rizzo@polito.it (R.G. Rizzo), andrea.calimera@polito.it (A. Calimera), zhouj@uestc.edu.cn (J. Zhou).

https://doi.org/10.1016/j.vlsi.2018.04.008
Received 21 December 2017; Received in revised form 18 March 2018; Accepted 20 April 2018
Available online XXX

Please cite this article in press as: R.G. Rizzo, et al., Approximate Error Detection-Correction for efficient Adaptive Voltage Over-Scaling, Integration, the VLSI Journal (2018), https://doi.org/10.1016/j.vlsi.2018.04.008

[5,7]. For error-resilient applications, one may just decide not to apply any correction [2]. Despite the impressive results achieved by Razor and Razor-based AVOS strategies, there's much room for improvement. That's the target of this paper.

1.2. Objectives

The broad objective of our work is to further improve existing error-driven AVOS schemes, both in terms of efficiency and flexibility. In pursuing this goal, we took into consideration the following requirements/constraints: (*i*) do not apply any "irreversible" modification of the circuit (i.e. re-synthesis stages); (*ii*) to be dynamic (i.e. adapt to different context and workloads), and tunable (i.e. support post-silicon calibration); (*iii*) to avoid dedicated optimization tools/algorithms that may be hard to integrate into industrial design kits.

The main intuition is that the availability of an Approximate Error Detection-Correction mechanism (*AED-C*, hereafter) represents a smart option to control the accuracy-energy tradeoff. The proposed AED-C scheme leverages the resolution of the error detection as a knob to dynamically accelerate (or slow-down) AVOS thus to achieve lower energy consumption (or high QoR).

1.3. Contributions

Our implementation of the AED-C mechanism passes through the use of (Razor-based) elastic timing sensors, namely, timing sensors with a tunable detection window (DW). Circuits implemented with our AED-C show a dual operating mode: (*i*) *Slow* AVOS mode, where DW is taken as large as possible, 50% of the clock period (T_{clk}), such that error detection is maximized; (*ii*) *Aggressive* AVOS mode, where DW is reduced in order to relax the error detection accuracy and to give AVOS the opportunity to maximize energy savings at the cost of QoR. The power-management unit may select the proper mode depending on the requirements imposed at the application level and/or other external variables, such as the remaining battery lifetime. Starting from the preliminary study reported in Ref. [9], this work reports a detailed discussion of the working principles of our AED-C technique. A parametric analysis conducted on a set of representative benchmarks (i.e. MAC, IIR and FIR Filters) empirically discloses its efficiency, also providing an assessment of the QoR-energy trade-off. The collected results show substantial energy savings w.r.t. a state-of-art Razor-based AVOS scheme: up to 44.7% on average for *Aggressive* mode and 10.9% for *Slow* mode. Moreover, the area overhead introduced by AED-C is much lower than that of Razor: 3.3% on average against 62.0%. As an additional piece of information, the analysis over a realistic error resilient application is reported: a Forward Discrete Cosine Transform Unit (FDCT) integrated into a JPEG compression system. For such test-case, simulations run on a set of images show average energy savings close to 52% (w.r.t. Baseline FDCT, i.e. w/o Razors) while guaranteeing a PSNR of 48.45 dB (w.r.t. Baseline JPEG pictures). These results get more appealing if one considers that the overhead introduced by a standard Razor strategy cannot be applied as it does not reach the timing closure of the design (more details available in the experimental section).

2. Beyond state-of-the-art

2.1. Existing Adaptive Voltage Over-Scaling approaches

Several embodiments of the AVOS principle are available in the literature [10]. They differ in terms of (*i*) the method used to detect/correct errors and/or (*ii*) the circuit optimization applied to mitigate/compensate the occurrence of errors. Despite such differences, they all show common characteristics that prevent their use for the implementation of a flexible AVOS strategy (the target of this work). In particular: (*i*) they alter the topology/structure of the circuit, thus loosing some of the good characteristics that usually bring to better voltage scaling profiles; (*ii*) they are static solutions, i.e. do not support an adaptive (on-line) tuning at post-silicon stage; (*iii*) they make use of additional logic blocks designed with custom synthesis procedures. The following text elaborates on these aspects for the most representative classes of AVOS schemes, emphasizing the main weakness addressed by our strategy.

Error Detection-Correction Schemes. Within this class, Razor [4,5] is the main reference. It is based on Razor-FFs that check the correctness of the circuit through a time-skewed comparison of the sampled values. The key idea of Razor is that of tuning the supply voltage by monitoring the error rate. This reduces the voltage margins and it helps to improve energy efficiency. Razor finds application in microprocessors architectures, where the error recovery is implemented by a three-stage mechanism: first, stall the pipeline; second, refresh flip-flops that experienced an error; third, re-execute the last pipe-cycle. Both error detection and correction are performed locally, i.e. on the Razor-FFs.

Razor II [7], an extension of Razor, avoids possible metastability issues by checking the occurrence of errors through a transition detector. Moreover, it shifts the error-recovery mechanism at a "software" level, that is, the FFs are in charge of error detection, while the correction is performed through a full replay of the latest instruction. This strategy significantly reduces the size the Razor-FF at the cost of overall performance penalties. The results collected on a 64-bit microprocessor show 33% average energy savings and an instruction-per-cycle degradation of 0.2%.

Authors in Refs. [11,12] proposed two alternative circuits for error detection and correction which prevent metastability; they replace the Razor FFs with time borrowing Latches. Silicon measurements show improved AVOS profiles with remarkable power reduction (37%).

Existing Razor-based strategies suffer from the so-called *short-path race*, which imposes the adoption of tedious hold-fixing procedures run at design time. Hold-fixing (when converging) may induce a heavy modification of the circuit's characteristic: paths are substantially shifted towards the clock edge. This issue (as proven by our simulation results) represents a serious impediment for effective use of AVOS.

Prediction based elastic-clocking. The authors of [13] propose a design paradigm, called CRISTA, that implements AVOS under a desired frequency constraint. The basic idea is to isolate the most critical paths through a custom re-synthesis stage that reshapes the original paths distribution. The Vdd is then tuned such that the most critical paths violate the set-up time, while the remaining non-critical paths run error-free. The activation of the critical paths is predicted by a dedicated control logic that works as a logic error sensor. The flag returned by this sensor triggers timing speculations: an extra clock-cycle is given when long-paths are excited. The CRISTA design methodology ensures activation rates of the long paths are low enough to avoid excessive performance penalties. For a two-stage pipelined ALU, CRISTA allows to reduce the power consumption by 40% with a mere 9% area overhead.

The solution presented in Refs. [14,15] is a variant of the CRISTA paradigm which still exploits the concept of variable latency units. Results show 45% power savings w.r.t. the baseline implementation. When tested on a 5-stage pipeline micro-architecture using SPEC2K benchmarks the throughput penalty is 4%.

No matter its actual implementation, CRISTA is applied at design time, namely, both the selection of the critical paths and the synthesis of the activation function (or variable latency units) are done statically using some a-priori knowledge of the circuit. However, process variations represent a serious concern; the path distribution may change due to manufacturing imperfections and some critical paths may result uncovered by the activation function (defined at design time). As far as we know, a practical solution to make CRISTA adaptive/tunable for post-silicon calibration does not exist; indeed, "guard-banding" (i.e., an over-selection of the critical paths that have to be isolated) seems the main option available. This exacerbates the design overhead. Also,