
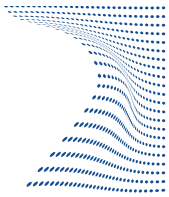

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (31.st cycle)

Energy-Accuracy Scaling in Digital ICs

Static and Adaptive Design Methods and Tools

Roberto Giorgio Rizzo
* * * * * *

Supervisors
Prof. E. Macii, Supervisor
Prof. A. Calimera, Co-supervisor

Doctoral Examination Committee:
Prof. Alberto Nannarelli, Referee, DTU - Danmarks Tekniske Universite
Prof. Michele Magno, Referee, Eidgenössische Technische Hochschule Zürich
Prof. Andrea Acquaviva, Università di Bologna
Prof. Silvia Chiusano, Politecnico di Torino
Prof. Paolo Garza, Politecnico di Torino

Politecnico di Torino
July 16, 2019

Summary

Energy efficiency has become the main constraint for most of today's information and communication technologies, from those involving high-performance computing (e.g., cloud services) to those deployed on low-power applications (e.g., portable systems for the Internet-of-Things). In the past decades, the pursuit of energy efficiency was mainly supported through the advance of the underlying CMOS technology. Moving towards a new node was the guarantee to achieve more than 90% of energy savings. However, as soon as the CMOS entered the nanometric regime, improvements brought by a technology shift have shrunk substantially, reaching 20% and then further less generation by generation. To make matters worse, production costs raised dramatically due to the technological impediments imposed by physical geometries below the 28 nm mark. This made technology scaling impractical for many cost-sensitive applications.

New sophisticated energy-aware design practices were then introduced to alleviate the suffering of a slow technology scaling. Very soon, low-power and energy-management techniques become the actual kernel of any design and optimization flow. Unfortunately, also design techniques are not fully renewable, namely, their effectiveness degrades with the advance of the technology nodes. This is the case of voltage scaling, for instance, which encountered the 1.0 V plateau that still holds today, but also other architectural-level techniques, such as multi-core/many-core solutions, which have been seriously limited by stringent dark-silicon constraints.

The end of Moore's law is not just a technology issue; it is also the prelude of a design crisis that will soon require to rethink the optimization and integration strategy of digital circuits and systems. A radical solution to all these concerns has to come yet. However, the recent growth of data-centric applications is opening to new design paradigms that alleviate the pressure. Much room is at the application-level indeed, where alternative energy-management knobs are available. The basic idea is that of integrating the quality-of-results as a new dimension in the design space. Leveraging the intrinsic error-resilience of data-centric applications, it is thereby possible to implement an *Energy-Accuracy Scaling* (EAS) which is orthogonal to the technology adopted and the low-power design strategy deployed. At the basis of this concept, there is the simple intuition that an application whose output can be degraded without affecting the quality perceived by the user may require lower energy consumption for the same amount of work.

The broad objective of this dissertation is to introduce advanced design solutions that improve the approach the EAS paradigm is implemented. Two new strategies are presented which reduce the design overhead of classical approximate solutions; according to the revisited taxonomy introduced in this thesis, one of the proposed strategies belongs to the class of *Adaptive* EAS, while the second falls under the label of *Static* EAS. With *Adaptive* EAS, the optimal energy-accuracy tradeoff is achieved by measuring some quality metrics directly on-chip, at run-time, establishing a feedback loop that drives the energy minimization. These metrics can be obtained by explicitly measuring the output accuracy, or by indirect measurements, e.g., through the output error rate. With *Static* EAS, the energy-accuracy tradeoff is fixed at design-time by functional speculation, i.e., a modification of the logic functionality through algorithmic or circuit simplifications which induce energy savings for a worst-case accuracy loss.

The Adaptive solution encompasses the extension of the conventional Error Detection-Correction techniques for data-driven voltage scaling in order to trade system accuracy for energy reduction. The new mechanism, called *Approximate Error Detection-Correction* (AED-C), is built upon in-situ *elastic timing monitors* which allow to implement a lightweight error management scheme.The AED-C implements EAS using the error detection coverage as a knob: a low error coverage accelerates supply voltage over-scaling thus to achieve more significant energy savings at the cost of quality-of-result; a high error coverage lessens the voltage scaling leading to higher accuracy at the cost of lower energy savings. As EAS does not have to ensure full error coverage, the traditionally large area/energy overhead of conventional techniques is drastically reduced. Simulations over a representative set of applications/circuits, e.g., Multiply-Accumulate (MAC) unit, Discrete Cosine Transform (DCT), FIR and IIR filters, provide a comparative analysis with the state-of-the-art techniques. The collected results show that AED-C substantially reduces the average energy-per-operation and the area overhead, still guaranteeing reasonable accuracy.

The static EAS strategy, instead, is developed exploiting Machine Learning theories which suggest alternative forms to represent relationships among data. Such theories find their application in the Boolean domain, where logic functions can be described as inference rules. The novel paradigm, named as *Inferential Logic*, leverages the concept of statistical inference for the design of combinational logic circuits that are able to mimic Boolean functions to a certain degree of accuracy. These inferential logic circuits run *quasi-exact* computations trading energy efficiency for accuracy in error-resilient applications. The figures-of-merit of an *Inferential Multiplier* are quantified using representative image processing applications as a case study. A comparative analysis against a state-of-the-art Booth Multiplier proves the inferential logic representation simplifies the circuit complexity reducing the overall area/delay. As a result, the inferential multiplier can exploit latency reduction for power optimization guaranteeing a fixed average accuracy.