PhD Thesis

Approximate@runtime: Playing with accuracy at run-time for low-power flexible circuits in IoT nodes

Keywords: approximate computing, inexact arithmetic, low-power multicore architectures, arithmetic operators, accuracy analysis, VLSI, FPGA

Laboratory: INRIA – CAIRN team <https://team.inria.fr/cairn>

Contact: Olivier Sentieys <olivier.sentieys@inria.fr>

Most traditional and emerging applications are able to produce acceptable results although relying on inexact or approximate computations. In addition to well-established applications (signal, image, vision, wireless communications, etc.), recently emerging applications such as machine learning, data mining and web search also exhibit the property to be inherently resilient to errors. Relying on application resilience, approximate computing has become a major field of research in the past few years. Approximate or inexact calculation provides energy gains by exploiting the tradeoff between energy and application quality. Again, the gain in energy between a low-precision 8-bit operation suitable for vision and a 64-bit double-precision floating point operation necessary for high-precision scientific computations, can reach up to 50x by considering storage, transport and computing of the data. By relaxing the need for fully precise or completely deterministic operations, approximate computing techniques allow substantially improved energy efficiency. There is strong research activity on these topics in the Cairn team.

Choosing the right computation precision at the right time during the execution, while preserving the application functionality in reasonable bounds, is another promising approach for improving significantly energy efficiency. Nevertheless, managing the quantity of performed error for a given application is an art. Turning this art into computer architecture innovations is a big challenge for making approximate computing a standard in efficient computing systems.

This thesis takes place in the ANR Artefact international project (CEA Leti, INRIA, INSA, EPFL, CSEM) whose aim is to leverage inexact and exact near-threshold and sub-threshold circuit design to achieve major energy consumption reductions by enabling adaptive accuracy control of applications. We propose to address, in a consistent fashion, the entire design stack, from physical hardware design, up to software application analysis, compiler optimizations, and dynamic energy management. We do believe that combining sub-near-threshold with inexact circuits on the hardware side and, in addition, extending this with intelligent and adaptive power management on the software side will produce outstanding results in terms of energy reduction, i.e., at least one order of magnitude, in IoT. INRIA CAIRN will contribute along two research directions: (1) approximate, ultra low-power circuit design and (2) accuracy-energy trade-offs in software.

As an example, traditionally, ultra-low-power radio transceivers are designed to operate at a fixed performance level. In practice, however, transmission quality varies with time not only because of fluctuations in received signal strength but also due to time-varying radio interference. Conceptually, therefore, a flexible transceiver able to autonomously scale its performance with respect to the instantaneous channel conditions and therefore only consume the minimum required amount of energy at any given time will achieve significant energy savings. And considering that 30 dB dynamic range fluctuations correspond to a linear factor of 1000, a channel-aware transceiver able to instantaneously scale its energy consumption by similarly important factors will lower its average power consumption and thus reach the 10-fold lifetime improvement goal. This contribution leverages our background on adaptive wireless receivers and on the analysis of variable levels of computation precision in wireless applications.

The contributions of thesis will cover some (but not all) of the following topics:

Accuracy Evaluation Our previous work on algorithmic-level accuracy evaluation due to approximation errors will be extended to other types of error such as those due to recent approximate operators or to soft errors. Recent and future technologies are more sensitive to transient faults and fault tolerance becomes one of the major challenges of system-on-chip design. Moreover, supply voltage reduction to decrease power consumption increases the probability of transient fault occurrence. These transient faults lead to erroneous output values. We will build a

framework for analyzing the robustness of the application to transient faults and to approximate operators in general.

Dealing with Errors for Low-Power Computing In usual arithmetic operators, all internal signals are computed to ensure exact or very accurate (w.r.t. the target precision) values for the results. This leads to accurate computations but high implementation cost and energy. There are specific arithmetic solutions where the internal quality of the arithmetic result is reduced to lower the power consumption. In estimated arithmetic for instance, some internal complex carry or signal schemes are over-simplified. This leads to computations with a reduced accuracy for some operands. Specific arithmetic operators have to be designed to ensure a good average accuracy as well as a very small probability to get totally wrong results. We plan to study arithmetic solutions with integrated internal approximation schemes (representations of numbers, arithmetic algorithms and dedicated tools). In a second time, we plan to develop tools to manage circuit designs based on such arithmetic solutions. We also plan to study methods to transform a numerical algorithm using a standard arithmetic support into an algorithm based on arithmetic datapath and/or operators with internal approximation schemes.

Hybrid Floating-Point and Fixed-Point Arithmetics Heterogeneous SoCs often integrate one or several cores supporting floating-point arithmetic while the others support only fixed-point arithmetic. Recent DSPs integrate floating-point and fixed-point units. The challenge is to hence find the good trade-off between the numerical accuracy, dynamic range and the implementation cost. In hardware implementation, custom floating-point data types can be considered by adjusting the word-length of the exponent and the mantissa part. Another objective we will explore is to integrate, in our methods and tools, models and optimized operators for floating-point arithmetic in order to design mixed systems integrating fixed-point and floating-point arithmetic with IEEE 754 or custom data types.

Approximate Compilers for Approximate Computing Matlab/Simulink is a popular framework for prototyping embedded applications, but is not suited for implementing the final system. These prototypes are therefore rewritten to enable their implementation. This rewritting often involve altering the initial algorithm to ease its mapping on the target. Examples of such modifications include floating point to fixed point arithmetic, but also algorithmic optimizations (relaxing the feedback loops, using sub-sampling, etc.) that help exploring coarser performance/accuracy trade-off. Theses transformations require significant application domain expertise (signal/image process- ing) and may raise design issues since they alter the semantics of the initial program. They are therefore not considered as relevant from a compiler optimization point of view. However, the need for low power design and quest for performance may soon force us to revise this principle. In this context, we want to study how some of these techniques can be automated through Scilab/Matlab source level transformations, as automation would enable a more systematic algorithmic design space exploration to application designers. For this research direction, we will build on the expertise of the group in both optimizing compiler and signal processing system implementation. Of course, for this toolbox to be usable in practice, users should be able to have quantitative estimates (or even guarantees) on the errors induced by these transformations and their impact on the whole program. Here again, we believe that the leading expertise of the group in accuracy analysis will help us address this challenge.

References

[Art16] ANR Artefact, projet international franco-suisse, http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-15-CE25-0015

[Eco] Oh, that's near enough computing: Letting microchips make a few mistakes here and there could make them much faster and more energy-efficient, The Economist, June 2012, http://www.economist.com/node/21556087

[Lin11] A. Lingamneni, C. Enz, J. L. Nagel, K. Palem, and C. Piguet. Energy parsimonious circuit design through probabilistic pruning. In IEEE/ACM Design, Automation Test in Europe Conference (DATE), pages 1–6, 2011.

[Din11] Florent de Dinechin and Bogdan Pasca. Designing custom arithmetic data paths with FloPoCo. IEEE Design & Test of Computers, 28(4):18--27, July 2011.

[Lu04] S.-L. Lu, "Speeding up processing with approximation circuits," Computer, vol. 37, no. 3, pp. 67-73, 2004.

[Kim11] J. Kim and S. Tiwari, "Inexact computing for ultra low-power nanometer digital circuit design," in IEEE/ACM Intl. Symp. on Nanoscale Architectures (NANOARCH), pp. 24-31, 2011.

[Men] D. Menard, R. Rocher, and O. Sentieys. Analytical Fixed-Point Accuracy Evaluation in Linear Time-Invariant Systems. IEEE Transactions on Circuits and Systems I: Regular Papers, 55(10):3197-3208, 2008.

[Par14] K. N. Parashar, D. Menard, and O. Sentieys. Accelerated Performance Evaluation of Fixed-Point Systems With Un-Smooth Operations. IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, 33(4):599–612, 2014.

[Dee14] G. Deest, T. Yuki, O. Sentieys, and S. Derrien. Toward Scalable Source Level Accuracy Analysis for Floating-point to Fixed-point Conversion. In IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2014.
[Bar16] B. Barrois, K. Parashar, and O. Sentieys. Leveraging Power Spectral Density for Scalable System-Level Accuracy Evaluation. In IEEE/ACM Conference on Design Automation and Test in Europe (DATE), page 6, Dresden, Germany, March 2016.
[Der08] S. Derrien, S. Rajopadhye, P. Quinton, and T. Risset. High-Level Synthesis of Loops Using the Polyhedral Model. In High-Level Synthesis : From Algorithm to Digital Circuit, pages 215–230. Spinger, 2008.
[Flo13] A.Floch et al. GeCoS: A framework for prototyping custom hardware design flows. In 13th IEEE International Conference on Source Code Analysis and Manipulation (SCAM), pages 100–105, Sept. 2013.
[Cla15] F. Cladera, M. Gautier, and O. Sentieys. Energy-Aware Computing via Adaptive Precision under Performance Constraints in OFDM Wireless Receivers. In IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pages 591 -- 596, Montpellier, France, July 2015. (Best Paper)
[Pas11] A. Pasha, S. Derrien, and O. Sentieys. System level synthesis for wireless sensor node controllers: A complete design flow. ACM Transactions on Design Automation of Electronic Systems (TODAES), 17(1):2.1-2.24, January 2011.