A Comprehensive Analysis of Approximate Computing Techniques: From Component- to Application-Level

Alberto Bosio, Daniel Menard, Olivier Sentieys

A Comprehensive Analysis of Approximate Computing Techniques: From Component- to Application-Level	
Organizer(s):	Alberto Bosio, École Centrale de Lyon - INL, FR
Reference person:	Alberto Bosio, École Centrale de Lyon - INL, FR
Speaker(s):	Alberto Bosio, École Centrale de Lyon - INL, FR Daniel Ménard, INSA Rennes - IETR, FR Olivier Sentieys, University of Rennes, INRIA/IRISA, FR
Preferred slot :AM vs PM	PM
DATE Track / Topic reference:	A7 Applications of Emerging Technologies T4 System-Level Dependability D9 Power Modeling, Low-Power Design, and Power-Aware Optimization

Motivations

Today, the concept of approximation is becoming more and more a “hot topic”. The relevance of this tutorial proposal is, therefore, very high, as it targets the design paradigm that has the potential to be the viable solution for dealing with the three main issues that are energy reduction, increase of performance and mitigating the intrinsic unreliability of the latest technology nodes.

In the last 5 years, numerous papers on approximate computing have been presented at DAC and DATE conferences. Several surveys on approximate computing have been published recently in different journals. This is the right moment to propose this tutorial which describes and classifies the different existing techniques and CAD frameworks.

Intended audience

This tutorial is primarily targeted at researchers and engineers working in the domain of Wireless Communications, Multimedia and in general, any Signal-Processing domain. System-level design engineers working on embedded software and hardware engineering aspects shall discover a new dimension, to improve performance and energy efficiency of their designs.

Objectives

The main objectives can be summarized as follows:

- Introduces the main concepts of Approximate Computing paradigm
- Define the different Approximate Techniques accordingly to the target layer (i.e., HW, SW)

- Present existing automated tools for the design of Approximate Computing Systems

Abstract

A new design paradigm, Approximate Computing (AxC), has been established to investigate how computing systems can be more energy efficient, faster, and less complex. Intuitively, instead of performing exact computation and, consequently, requiring a high amount of resources, AxC aims to selectively relax the specifications, trading accuracy off for efficiency. It has been demonstrated in the literature the effectiveness of imprecise computation for both software and hardware components implementing inexact algorithms, showing an inherent resiliency to errors.

This tutorial introduces basic and advanced topics on AxC. We intend to follow a bottom-up approach: from component, up to application-level. More in detail, we will first present existing approximate computing techniques according to three levels: hardware, data and computation.
At hardware level, functional approximation through inexact operators and voltage over-scaling will be detailed. At data level, approximation can be carried-out by using efficient arithmetic, precision scaling, less data or less-up-to-date data. We will present some compile-time results in terms of energy-efficiency, area, performance versus accuracy of computations when using customized arithmetic (fixed-point, floating-point) and also, we will try to derive some conclusions by comparing the different paradigms. At computation level, algorithmic transformations are used to reduce complexity by skipping or approximating parts of the processing. The concepts of loop perforation, early termination, memoïzation, and computation approximation will be detailed.

The last part of the tutorial is dedicated to methods and tools to exploit efficiently approximate computing. First of all, the different approaches to analyze approximation effects on application quality will be described. Then the complex problem of word-length optimization for fixed-point and floating-point is considered. Finally, the different frameworks for design space exploration will be detailed.

Necessary background

Knowledge of Computer Architecture and Computer Arithmetic

References

M. Barbareschi, A. Mazzeo, A. Bosio, “Design Exploration Tool for Approximate Algorithms (IDEA)”, http://wpage.unina.it/mario.barbareschi/iideaa

G.Rodrigues, F.Kastensmidt, V.Pouget, A.Bosio, “Performances VS Reliability: Approximate Computing can make the difference”, IOLTS’2018

B. Barrois, O. Sentieys, D. Menard. 2017. The hidden cost of functional approximation against careful data sizing: a case study. In *Proceedings of the Conference on Design, Automation & Test in*

Europe (DATE '17). Lausanne, 2017, pp. 181-186.

R. Ragavan, B. Barrois, C. Killian and O. Sentieys, "Pushing the limits of voltage over-scaling for error-resilient applications," *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, Lausanne, 2017, pp. 476-481.

E. Nogues, D. Menard and M. Pelcat, "Algorithmic-level Approximate Computing Applied to Energy Efficient HEVC Decoding," in *IEEE Transactions on Emerging Topics in Computing*.

Has the same tutorial (or a similar one) been **presented to other events** (if yes, list when/where)?

Yes. A similar tutorial will be presented by the same speakers at ESWeek 20018 in September 2018

Has the same **organizer proposed other tutorials** (if yes, list when/where and on what topic)?

Daniel Ménard and Olivier Sentieys have presented 3 tutorials (DATE 2014, DATE 2015, HiPeac 2016) on fixed-point refinement:

D. Menard, D. Novo, and K. Parashar, O. Sentieys. Fixed-point refinement, a guaranteed approach towards energy efficient computing, January 2016. Tutorial at HiPeac Conference.

O. Sentieys, D. Menard, D. Novo, and K.Parashar. Fixed-point refinement, a guaranteed approach towards energy efficient computing Tutorial at IEEE/ACM Design Automation and Test in Europe (DATE'15), Grenoble, March 2015.

D. Menard, D. Novo, K. Parashar, O. Sentieys,. Automatic fixed-point conversion : a gate way to high-level power optimization Tutorial at IEEE/ACM Design Automation and Test in Europe (DATE'14), Dresden, March 24 2014.

Tutorial material

Presentation slides will be available for download by the attendees

Tutorial plan

1. General introduction Motivations (30 min)
    - 1.1. Definition(s)
    - 1.2. Overview and classification of techniques
    - 1.3. Metrics
    - 1.4. Approaches

2. Techniques for approximate computing
    - 2.1. Data level approximation (30 min)
        - 2.1.1. Data representation - arithmetic
        - 2.1.2. Adaptive Precision scaling
        - 2.1.3. Less data

2.1.4. Less up-to-date data

2.2. Hardware level approximation (30 min)

2.2.1. Exact integer and floating-point operators

2.2.2. Inexact operators

2.2.3. Voltage over-scaling / Overclocking

2.2.4. Approximate memories

2.2.5. Comparison of fixed-point vs. inexact operators vs. custom floating-point

2.3. Computation level approximation (30 min)

2.3.1. Computation skip (Fine grained, coarse grained)

2.3.2. Computation approximation (Algorithm selection, Memoization, CNN)

3. Methods and tools for approximate computing

3.1. Analysis of approximation effect on application quality (30 min)

3.1.1. Error metrics

3.1.2. Approximate error modeling

3.1.3. Simulation-based approaches

3.1.4. Analytical approaches

3.2. Word-length optimization for fixed-point and floating-point (30mn)

3.2.1. Problem statement

3.2.2. State of the art algorithms

3.2.3. Case study of WLO

3.3. Design space exploration - (30 min)

3.3.1. Problem statement

3.3.2. State of the art (ACCEPT, REACT, PRECIMONIUS)

3.3.3. Issues

3.3.4. IIDEA Framework

3.3.5. Demo

3.3.6. Experimental Results

Biographies

Alberto Bosio received the PhD in Computer Engineering from the Politecnico di Torino, Italy in 2006. From 2007 to 2018 was an Associate Professor at LIRMM - University of Montpellier in France. He is now a Full Professor at the INL - Ecole Centrale de Lyon (France). His research interests include Approximate Computing, In-Memory Computing, Test and Diagnosis of Digital circuits and systems and Reliability. He co-authored 1 book, 3 patents 35 journals, and over 120 conference papers. He will be the chair of the ETTTC from January 2018. He is a member of the IEEE.
Web: http://www.lirmm.fr/~bosio/home/

Daniel Menard received the Ph.D. and HDR degrees in Signal Processing and Telecommunications from the University of Rennes, respectively in 2002 and 2011. From 2003 to 2012 he was Associate Professor at University of Rennes in France. He is currently Full Professor at INSA Rennes. His research activities focus on the energy efficient implementation of signal and image processing applications in embedded systems. His research topics include approximate computing, accuracy evaluation, fixed-point arithmetic, energy optimization in MPSoC, low power HEVC video encoding and decoding and embedded stereo-vision. He has published more than 100 papers in international journals and in international conferences. He is member of the DISPS Technical Committee of the IEEE Signal Processing Society.
Web: http://dmenard.perso.insa-rennes.fr

Olivier Sentieys is a Professor at the University of Rennes and holds an Inria Research Chair on Energy-Efficient Computing Systems. He has more than 20 years of expertise in the fields of system-on-chip architectures, reconfigurable systems and their associated CAD tools, finite arithmetic effects, numerical accuracy analysis and low-power sensor networks. He authored or coauthored more than 150 journal publications or peer-reviewed conference papers and hold 6 patents. In particular, his research on methods for analytical analysis of errors in reduced-precision arithmetic and word-length optimization since 2000 with more than 50 publications, can be considered as a pioneering work in the field of approximate computing.
Web: http://people.rennes.inria.fr/Olivier.Sentieys/