Thesis Project: Application-Level Tuning of Accuracy

Olivier Sentieys and Tomofumi Yuki
CAIRN, Inria - Rennes

Keywords: Approximate Computing, Accuracy Analysis, Run-time Adaptation.

Context: Energy consumption is one of the major issues in computing today shared by all domains in computer science, from high-performance computing to embedded systems. The two main factors that influence energy consumption is the execution time and data volume. Execution time directly impacts energy, as energy is the product of time and (average) power. Another large source of energy consumption is data transfer - the volume of data is directly proportional to the amount of energy consumed.

In the recent years, approximation is receiving renewed interests to improve both speed and energy consumption in embedded systems [1, 2, 3]. Many applications in embedded systems do not require high precision/accuracy, and hardware designers often seek for a good balance between accuracy, speed, energy, and area cost. Various techniques for approximate computing augment the design space by providing another set of design knobs for performance-accuracy trade-off.

This project is about developing methods for *systematic* exploration of the design space, including performance/accuracy modeling and (semi-)automation of designs. We target emerging System on Chip platforms (Xilinx Zynq or Intel SoC) that feature FPGAs tightly coupled with embedded processors. This class of platforms expose more complex and interesting design space where the trade-offs can take place at multiple compute resources, and communication layers.

In terms of applications, we are interested in those that have multiple candidate kernels for approximation. The various factors that influence accuracy is expected to have complex interplay when compositions of smaller kernels are jointly considered. Where to perform the trade-off becomes and important decision step in such contexts, and we seek to develop approaches to guide the design choices.

Subject: The high-level questions we pose ourselves are as follows:

- How to reason about the influence to accuracy at the system-level? Applications are composed of multiple kernels that can each be approximated. Is it better to heavily approximate a single kernel? Should we slightly approximate each kernel? The answers to these questions are not obvious, both from accuracy and speed point of view.
- How to control the degree of approximation? Can we dynamically tune the degree at run-time? The impact of accuracy-altering transformations are highly dependent on the input data. This motivates the need for run-time adaptation techniques that adapt to inputs at run-time.
- How does approximations interact with performance-oriented optimizations? Approximations can also be viewed as semantic non-preserving transformations that could be used as enabler transformations for other optimizations, e.g., relaxing dependences for increased parallelism.

We expect arithmetic precision to play an important role for energy savings due to its impact on data size. The interaction of the precision with other accuracy influencing features is one of the key problems that we would like to understand through this work.

Skills and Expectations: The student is expected to develop techniques for mathematical formulation of accuracy trade-off, and static/dynamic approaches for exposing accuracy as an additional design knob. We also expect to have prototype implementations of the developed techniques. The FPGA designs will primarily be done through High-Level Synthesis tools.

Desired skills include:

- Basic knowledge in linear algebra.
- Familiarity with the C language.
- Familiarity with FPGA design and/or HLS.

Mostly importantly, we seek highly motivated and active students.

References

[1] Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications*, 180(12):2526–2533, 2009.

[2] Benjamin Barrois, Olivier Sentieys, and Daniel Menard. The Hidden Cost of Functional Approximation Against Careful Data Sizing – A Case Study. In *Design, Automation & Test in Europe Conference & Exhibition (DATE 2017)*, Lausanne, France, 2017.

[3] Rengarajan Ragavan, Benjamin Barrois, Cedric Killian, and Olivier Sentieys. Pushing the Limits of Voltage Over-Scaling for Error-Resilient Applications. In *Design, Automation & Test in Europe Conference & Exhibition (DATE 2017)*, Lausanne, Switzerland, March 2017.