UMR IRISA

Activity Report 2019

Team CAIRN

Energy Efficient Computing Architectures

Joint team with Inria Rennes – Bretagne Atlantique

D3 – Architecture

Table of contents

Project-Team CAIRN

Creation of the Project-Team: 2009 January 01

Keywords:

Computer Science and Digital Science:

A1.1. - Architectures
A1.1.1. - Multicore, Manycore
A1.1.2. - Hardware accelerators (GPGPU, FPGA, etc.)
A1.1.8. - Security of architectures
A1.1.9. - Fault tolerant systems
A1.1.10. - Reconfigurable architectures
A1.1.12. - Non-conventional architectures
A1.2.5. - Internet of things
A1.2.6. - Sensor networks
A2.2. - Compilation
A2.2.1. - Static analysis
A2.2.4. - Parallel architectures
A2.2.6. - GPGPU, FPGA...
A2.2.7. - Adaptive compilation
A4.4. - Security of equipment and software
A8.10. - Computer arithmetic

Other Research Topics and Application Domains:

B4.5. - Energy consumption
B4.5.1. - Green computing
B4.5.2. - Embedded sensors consumption
B6.2.2. - Radio technology
B6.2.4. - Optic technology
B6.6. - Embedded systems
B8.1. - Smart building/home
B8.1.1. - Energy for smart buildings
B8.1.2. - Sensor networks for smart buildings

1. Team, Visitors, External Collaborators

Research Scientists

François Charot [Inria, Researcher]
Silviu-Ioan Filip [Inria, Researcher]
Tomofumi Yuki [Inria, Researcher]

Faculty Members

Olivier Sentieys [Team leader, Professor, Univ. Rennes, Inria Chair, HDR]
Emmanuel Casseau [Professor, Univ. Rennes, ENSSAT, Lannion, HDR]
Daniel Chillet [Professor, Univ. Rennes, ENSSAT, Lannion, HDR]
Steven Derrien [Professor, Univ. Rennes, ISTIC, Rennes, HDR]

Cédric Killian [Associate Professor, Univ. Rennes, IUT, Lannion]
Angeliki Kritikakou [Associate Professor, Univ. Rennes, ISTIC, Rennes]
Patrice Quinton [Ecole Normale Supérieure de Rennes, Emeritus, Rennes]
Christophe Wolinski [Professor, Univ. Rennes, Rennes, HDR]

Post-Doctoral Fellows
Mansureh Shahraki Moghaddam [Inria, Rennes, until May 2019]
Lei Mo [Inria, Rennes, until Apr 2019]

PhD Students
Joel Ortiz Sosa [Inria, Lannion, from Oct. 2016]
Nicolas Roux [Inria, granted by Brittany Region/LTC, Lannion, from Oct. 2016]
Mael Gueguen [Univ. Rennes, MENRT grant, Rennes, from Nov. 2016]
Minh Thanh Cong [Univ. Rennes, granted by USTH, Rennes, from May 2017]
Thibaut Marty [Univ. Rennes, granted by H2020 ARGO and Brittany Region, Rennes, from Sep. 2017]
Petr Dobias [Univ. Rennes, MENRT grant, Lannion, from Oct. 2017]
Van-Phu Ha [Inria, granted by ANR Artefact, Rennes, from Nov. 2017]
Romain Mercier [Inria, granted by DGA and Inria, Lannion, from Oct. 2018]
Minyu Cui [Univ. Rennes, granted by CSC-ENS, Rennes, from Sep. 2018]
Jaechul Lee [Univ. Rennes, granted by Brittany Region/LTC, Lannion, from Dec. 2018]
Davide Pala [Inria, granted by IPL ZEP, Rennes, from Jan. 2018]
Joseph Paturel [Inria, granted by RAPID FLODAM, Rennes, from Sep. 2018]
Thibault Allenet [Univ. Rennes, granted by CEA LIST, Saclay, from March 2019]
Corentin Ferry [Univ. Rennes, granted by ENS, Rennes, from Sep. 2019]
Adrien Gaonac'h [Univ. Rennes, granted by CEA LIST, Saclay, from Oct. 2019]
Genevieve Ndour [Univ. Rennes, granted by CEA Leti, Grenoble, from May 2016]
Audrey Lucas [CNRS, granted by DGA-PEC, Lannion, until Jan. 2019]

Technical staff
Mickaël Le Gentil [Research Engineer (half time), Univ. Rennes, Lannion]
Justine Bonnot [Inria, Engineer, from Nov. 2019]
Pierre Hallé [Inria, Engineer, Lannion]
Simon Rokicki [Inria, Engineer, Rennes]

Administrative Assistants
Nadia Derouault [Assistant, Inria, Rennes]
Emilie Carquin [Assistant, Univ. Rennes, ENSSAT, Lannion]

Visiting Scientists
Bernard Goossens [Professor, Univ. Perpignan, July 2019]
Sharad Sinha [Professor, IIT Goa, India, July 2019]

2. Overall Objectives

2.1. Overall Objectives

Abstract — The CAIRN project-team researches new architectures, algorithms and design methods for flexible, secure, fault-tolerant, and energy-efficient domain-specific system-on-chip (SoC). As performance and energy-efficiency requirements of SoCs, especially in the context of multi-core architectures, are continuously increasing, it becomes difficult for computing architectures to rely only on programmable processors solutions. To address this issue, we promote/advocate the use of reconfigurable hardware, i.e., hardware structures whose organization may change before or even during execution. Such reconfigurable chips offer high performance at a low energy cost, while preserving a high level of flexibility. The group studies these systems from three angles: (i) The invention and design of new reconfigurable architectures with an emphasis on flexible arithmetic operator design, dynamic reconfiguration management and low-power consumption. (ii) The development of their corresponding design flows (compilation and synthesis tools) to enable their automatic design from high-level specifications. (iii) The interaction between algorithms and architectures especially for our main application domains (wireless communications, wireless sensor networks and digital security).

Keywords — Architectures: Embedded Systems, System-on-Chip, Reconfigurable Architectures, Hardware Accelerators, Low-Power, Computer Arithmetic, Secure Hardware, Fault Tolerance. **Compilation and synthesis:** High-Level Synthesis, CAD Methods, Numerical Accuracy Analysis, Fixed-Point Arithmetic, Polyhedral Model, Constraint Programming, Source-to-Source Transformations, Domain-Specific Optimizing Compilers, Automatic Parallelization. **Applications:** Wireless (Body) Sensor Networks, High-Rate Optical Communications, Wireless Communications, Applied Cryptography.

The scientific goal of the CAIRN group is to research new hardware architectures for domain-specific SoCs, along with their associated design and compilation flows. We particularly focus on on-chip integration of specialized and reconfigurable accelerators. Reconfigurable architectures, whose hardware structure may be adjusted before or even during execution, originate from the possibilities opened up by Field Programmable Gate Arrays (FPGA) [50] and then by Coarse-Grain Reconfigurable Arrays (CGRA) [53], [65] [1]. Recent evolutions in technology and modern hardware systems confirm that reconfigurable systems are increasingly used in recent and future applications (see e.g. Intel/Altera or Xilinx/Zynq solutions). This architectural model has received a lot of attention in academia over the last two decades [56], and is now considered for industrial use in many application domains. One first reason is that the rapidly changing standards or applications require frequent device modifications. In many cases, software updates are not sufficient to keep devices on the market, while hardware redesigns remain too expensive. Second, the need to adapt the system to changing environments (e.g., wireless channel, harvested energy) is another incentive to use runtime dynamic reconfiguration. Moreover, with technologies at 28 nm and below, manufacturing problems strongly impact electrical parameters of transistors, and transient errors caused by particles or radiations also often appear during execution: error detection and correction mechanisms or autonomic self-control can benefit from reconfiguration capabilities.

As chip density increased, power or energy efficiency has become “the Grail” of all chip architects. With the end of Dennard scaling [60], multicore architectures are hitting the *utilisation wall* and the percentage of transistors in a chip that can switch at full frequency drops at a fast pace [54]. However, this unused portion of a chip also opens up new opportunities for computer architecture innovations. Building specialized processors or hardware accelerators can come with orders-of-magnitude gains in energy efficiency. Since from the beginning of CAIRN in 2009, we have been advocating heterogeneous multicores, in which general-purpose processors (GPPs) are integrated with specialized accelerators, especially when built on reconfigurable hardware, which provides the best trade-off between power, performance, cost and flexibility. Time has confirmed the importance of heterogeneous manycore architectures, which are prevalent today.

> Standard multicore architectures enable flexible software on fixed hardware, whereas reconfigurable architectures make possible **flexible software on flexible hardware**.

However, designing reconfigurable systems poses several challenges: the definition of the architecture structure itself, along with its dynamic reconfiguration capabilities, and its corresponding compilation or synthesis tools. The scientific goal of CAIRN is to tackle these challenges, leveraging the background and past experience of the team members. We propose to approach energy efficient reconfigurable architectures from three angles: (i) the invention and the design of new reconfigurable architectures or hardware accelerators, (ii) the development of their corresponding compilers and design methods, and (iii) the exploration of the interaction between applications and architectures.

3. Research Program

3.1. Panorama

The development of complex applications is traditionally split in three stages: a theoretical study of the algorithms, an analysis of the target architecture and the implementation. When facing new emerging applications such as high-performance, low-power and low-cost mobile communication systems or smart sensor-based systems, it is mandatory to strengthen the design flow by a joint study of both algorithmic and architectural issues.

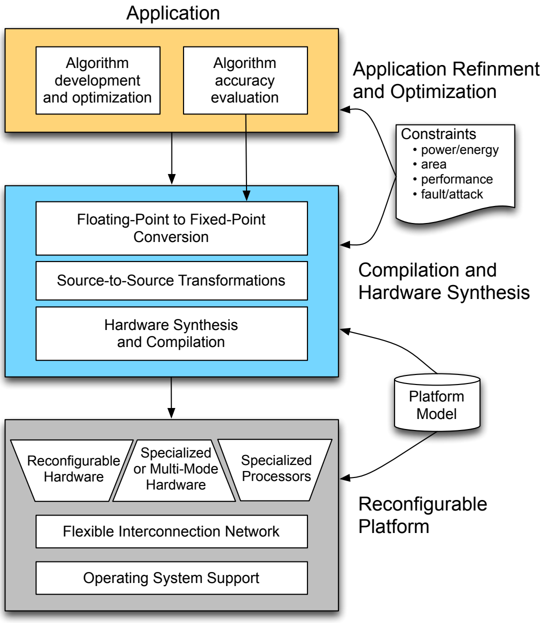
Application
Algorithm development and optimization
Algorithm accuracy evaluation
Application Refinment and Optimization
Constraints
• power/energy
• area
• performance
• fault/attack
Floating-Point to Fixed-Point Conversion
Source-to-Source Transformations
Hardware Synthesis and Compilation
Compilation and Hardware Synthesis
Platform Model
Reconfigurable Hardware
Specialized or Multi-Mode Hardware
Specialized Processors
Flexible Interconnection Network
Operating System Support
Reconfigurable Platform

Figure 1. CAIRN*'s general design flow and related research themes*

Figure 1 shows the global design flow we propose to develop. This flow is organized in levels corresponding to our three research themes: application optimization (new algorithms, fixed-point arithmetic, advanced representations of numbers), architecture optimization (reconfigurable and specialized hardware, application-specific processors, arithmetic operators and functions), and stepwise refinement and code generation (code transformations, hardware synthesis, compilation).
In the rest of this part, we briefly describe the challenges concerning **new reconfigurable platforms** in Section 3.2 and the issues on **compiler and synthesis tools** related to these platforms in Section 3.3.

3.2. Reconfigurable Architecture Design

Nowadays, FPGAs are not only suited for application specific algorithms, but also considered as fully-featured computing platforms, thanks to their ability to accelerate massively parallelizable algorithms much faster than their processor counterparts [68]. They can also be reconfigured dynamically. At runtime, partially reconfigurable regions of the logic fabric can be reconfigured to implement a different task, which allows for a better resource usage and adaptation to the environment. Dynamically reconfigurable hardware can also cope with hardware errors by relocating some of its functionalities to another, sane, part of the logic fabric. It could also provide support for a multi-tasked computation flow where hardware tasks are loaded on-demand at runtime. Nevertheless, current design flows of FPGA vendors are still limited by the use of one partial bitstream for each reconfigurable region and for each design. These regions are defined at design time and it is not possible to use only one bitstream for multiple reconfigurable regions nor multiple chips. The multiplicity of such bitstreams leads to a significant increase in memory. Recent research has been conducted in the domain of task relocation on a reconfigurable fabric. All related work has been conducted on architectures from commercial vendors (e.g., Xilinx, Altera) which share the same limitations: the inner details of the bitstream are not publicly known, which limits applicability of the techniques. To circumvent this issue, most dynamic reconfiguration techniques are either generating multiple bitstreams for each location [52] or implementing an online filter to relocate the tasks [62]. Both of these techniques still suffer from memory footprint and from the online complexity of task relocation.

Increasing the level and grain of reconfiguration is a solution to counterbalance the FPGA penalties. Coarse-grained reconfigurable architectures (CGRA) provide operator-level configurable functional blocks and word-level datapaths [69], [57], [67]. Compared to FPGA, they benefit from a massive reduction in configuration memory and configuration delay, as well as for routing and placement complexity. This in turns results in an improvement in the computation volume over energy cost ratio, although with a loss of flexibility compared to bit-level operations. Such constraints have been taken into account in the design of DART[9], Adres [65] or polymorphous computing fabrics[11]. These works have led to commercial products such as the PACT/XPP [51] or Montium from Recore systems, without however a real commercial success yet. Emerging platforms like Xilinx/Zynq or Intel/Altera are about to change the game.

In the context of emerging heterogenous multicore architecture, CAIRN advocates for associating general-purpose processors (GPP), flexible network-on-chip and coarse-grain or fine-grain dynamically reconfigurable accelerators. We leverage our skills on microarchitecture, reconfigurable computing, arithmetic, and low-power design, to discover and design such architectures with a focus on: reduced energy per operation; improved application performance through acceleration; hardware flexibility and self-adaptive behavior; tolerance to faults, computing errors, and process variation; protections against side channel attacks; limited silicon area overhead.

3.3. Compilation and Synthesis for Reconfigurable Platforms

In spite of their advantages, reconfigurable architectures, and more generally hardware accelerators, lack efficient and standardized compilation and design tools. As of today, this still makes the technology impractical for large-scale industrial use. Generating and optimizing the mapping from high-level specifications to reconfigurable hardware platforms are therefore key research issues, which have received considerable interest over the last years [55], [70], [66], [64], [63]. In the meantime, the complexity (and heterogeneity) of these platforms has also been increasing quite significantly, with complex heterogeneous multi-cores architectures becoming a *de facto* standard. As a consequence, the focus of designers is now geared toward optimizing overall system-level performance and efficiency [61]. Here again, existing tools are not well suited, as they fail at providing a unified programming view of the programmable and/or reconfigurable components implemented on the platform.

In this context, we have been pursuing our efforts to propose tools whose design principles are based on a tight coupling between the compiler and the target hardware architectures. We build on the expertise of the team members in High Level Synthesis (HLS) [5], ASIP optimizing compilers [12] and automatic parallelization for massively parallel specialized circuits [2]. We first study how to increase the efficiency of standard programmable processors by extending their instruction set to speed-up compute intensive kernels. Our focus is on efficient and exact algorithms for the identification, selection and scheduling of such instructions [6]. We address compilation challenges by borrowing techniques from high-level synthesis, optimizing compilers and automatic parallelization, especially when dealing with nested loop kernels. In addition, and independently of the scientific challenges mentioned above, proposing such flows also poses significant software engineering issues. As a consequence, we also study how leading edge software engineering techniques (Model Driven

Engineering) can help the Computer Aided Design (CAD) and optimizing compiler communities prototyping new research ideas [4].

Efficient implementation of multimedia and signal processing applications (in software for DSP cores or as special-purpose hardware) often requires, for reasons related to cost, power consumption or silicon area constraints, the use of fixed-point arithmetic, whereas the algorithms are usually specified in floating-point arithmetic. Unfortunately, fixed-point conversion is very challenging and time-consuming, typically demanding up to 50% of the total design or implementation time. Thus, tools are required to automate this conversion. For hardware or software implementation, the aim is to optimize the fixed-point specification. The implementation cost is minimized under a numerical accuracy or an application performance constraint. For DSP-software implementation, methodologies have been proposed [7] to achieve fixed-point conversion. For hardware implementation, the best results are obtained when the word-length optimization process is coupled with the high-level synthesis [58]. Evaluating the effects of finite precision is one of the major and often the most time consuming step while performing fixed-point refinement. Indeed, in the word-length optimization process, the numerical accuracy is evaluated as soon as a new word-length is tested, thus, several times per iteration of the optimization process. Classical approaches are based on fixed-point simulations [59]. Leading to long evaluation times, they can hardly be used to explore the design space. Therefore, our aim is to propose closed-form expressions of errors due to fixed-point approximations that are used by a fast analytical framework for accuracy evaluation [10].

4. Application Domains

4.1. Panorama

keywords: Wireless (Body) Sensor Networks, High-Rate Optical Communications, Wireless Communications, Applied Cryptography, Machine Learning, Deep Learning, Image ans Signal Processing.

Our research is based on realistic applications, in order to both discover the main needs created by these applications and to invent realistic and interesting solutions.

Wireless Communication is our privileged application domain. Our research includes the prototyping of (subsets of) such applications on reconfigurable and programmable platforms. For this application domain, the high computational complexity of the 5G Wireless Communication Systems calls for the design of high-performance and energy-efficient architectures. In **Wireless Sensor Networks** (WSN), where each wireless node is expected to operate without battery replacement for significant periods of time, energy consumption is the most important constraint. Sensor networks are a very dynamic domain of research due, on the one hand, to the opportunity to develop innovative applications that are linked to a specific environment, and on the other hand to the challenge of designing totally autonomous communicating objects.

Other important fields are also considered: hardware cryptographic and security modules, high-rate optical communications, machine learning, data mining, and multimedia processing.

5. New Software and Platforms

5.1. Gecos

Generic Compiler Suite

KEYWORDS: Source-to-source compiler - Model-driven software engineering - Retargetable compilation

SCIENTIFIC DESCRIPTION: The Gecos (Generic Compiler Suite) project is a source-to-source compiler infrastructure developed in the Cairn group since 2004. It was designed to enable fast prototyping of program analysis and transformation for hardware synthesis and retargetable compilation domains.

Gecos is Java based and takes advantage of modern model driven software engineering practices. It uses the Eclipse Modeling Framework (EMF) as an underlying infrastructure and takes benefits of its features to make it easily extensible. Gecos is open-source and is hosted on the Inria gforge.

The Gecos infrastructure is still under very active development, and serves as a backbone infrastructure to projects of the group. Part of the framework is jointly developed with Colorado State University and between 2012 and 2015 it was used in the context of the FP7 ALMA European project. The Gecos infrastructure is currently used by the EMMTRIX start-up, a spin-off from the ALMA project which aims at commercializing the results of the project, and in the context of the H2020 ARGO European project.

FUNCTIONAL DESCRIPTION: GeCoS provides a programme transformation toolbox facilitating parallelisation of applications for heterogeneous multiprocessor embedded platforms. In addition to targeting programmable processors, GeCoS can regenerate optimised code for High Level Synthesis tools.

- Participants: Tomofumi Yuki, Thomas Lefeuvre, Imèn Fassi, Mickael Dardaillon, Ali Hassan El Moussawi and Steven Derrien
- Partner: Université de Rennes 1
- Contact: Steven Derrien
- URL: http://gecos.gforge.inria.fr

5.2. ID-Fix

Infrastructure for the Design of Fixed-point systems

KEYWORDS: Energy efficiency - Dynamic range evaluation - Accuracy optimization - Fixed-point arithmetic - Analytic Evaluation - Embedded systems - Code optimisation

SCIENTIFIC DESCRIPTION: The different techniques proposed by the team for fixed-point conversion are implemented on the ID.Fix infrastructure. The application is described with a C code using floating-point data types and different pragmas, used to specify parameters (dynamic, input/output word-length, delay operations) for the fixed-point conversion. This tool determines and optimizes the fixed-point specification and then, generates a C code using fixed-point data types (ac_fixed) from Mentor Graphics. The infrastructure is made-up of two main modules corresponding to the fixed-point conversion (ID.Fix-Conv) and the accuracy evaluation (ID.Fix-Eval).

FUNCTIONAL DESCRIPTION: ID.Fix focuses on computational precision accuracy and can provide an optimised specification using fixed point arithmetic from a C source code with floating point data types. Fixed point arithmetic is very widely used in embedded systems as it provides better performance and is much more energy efficient. ID.Fix used an analytic programme model which means it can explore more solutions and thereby produce much more efficient code.

- Participant: Olivier Sentieys
- Partner: Université de Rennes 1
- Contact: Olivier Sentieys
- URL: http://idfix.gforge.inria.fr

5.3. SmartSense

KEYWORDS: Wireless Sensor Networks - Smart building - Non-Intrusive Appliance Load Monitoring

FUNCTIONAL DESCRIPTION: To measure energy consumption by equipment in a building, NILM techniques (Non-Intrusive Appliance Load Monitoring) are based on observation of overall variations in electrical voltage. This avoids having to deploy watt-meters on every device and thus reduces the cost. SmartSense goes a step further to improve on these techniques by combining sensors (light, temperature, electromagnetic wave, vibration and sound sensors, etc.) to provide additional information on the activity of equipment and people. Low-cost sensors can be energy-autonomous too.

- Contact: Olivier Sentieys

5.4. Platforms

5.4.1. *Zyggie: a Wireless Body Sensor Network Platform*

KEYWORDS: Health - Biomechanics - Wireless body sensor networks - Low power - Gesture recognition - Hardware platform - Software platform - Localization

SCIENTIFIC DESCRIPTION: Zyggie is a hardware and software wireless body sensor network platform. Each sensor node, attached to different parts of the human body, contains inertial sensors (IMU) (accelerometer, gyrometer, compass and barometer), an embedded processor and a low-power radio module to communicate data to a coordinator node connected to a computer, tablet or smartphone. One of the system's key innovations is that it collects data from sensors as well as on distances estimated from the power of the radio signal received to make the 3D location of the nodes more precise and thus prevent IMU sensor drift and power consumption overhead. Zyggie can be used to determine posture or gestures and mainly has applications in sport, healthcare and the multimedia industry.

FUNCTIONAL DESCRIPTION: The Zyggie sensor platform was developed to create an autonomous Wireless Body Sensor Network (WBSN) with the capabilities of monitoring body movements. The Zyggie platform is part of the BoWI project funded by CominLabs. Zyggie is composed of a processor, a radio transceiver and different sensors including an Inertial Measurement Unit (IMU) with 3-axis accelerometer, gyrometer, and magnetometer. Zyggie is used for evaluating data fusion algorithms, low power computing algorithms, wireless protocols, and body channel characterization in the BoWI project.

The Zyggie V2 prototype (see Figure 2) includes the following features: a 32-bit micro-controller to manage a custom MAC layer and process quaternions based on IMU measures, and an UWB radio from DecaWave to measure distances between nodes with Time of Flight (ToF).

- Participants: Arnaud Carer and Olivier Sentieys
- Partners: Lab-STICC, Université de Rennes 1
- Contact: Olivier Sentieys
- URL: https://bowi.cominlabs.u-bretagneloire.fr/zyggie-wbsn-platform

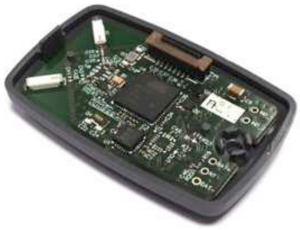

Figure 2. CAIRN's Zyggie platform for WBSN

5.4.2. *E-methodHW: an automatic tool for the evaluation of polynomial and rational function approximations*

KEYWORDS: function approximation, FPGA hardware implementation generator

SCIENTIFIC DESCRIPTION: E-methodHW is an open source C/C++ prototype tool written to exemplify what kind of numerical function approximations can be developed using a digit recurrence evaluation scheme for polynomials and rational functions.

FUNCTIONAL DESCRIPTION: E-methodHW provides a complete design flow from choice of mathematical function operator up to optimised VHDL code that can be readily deployed on an FPGA. The use of the E-method allows the user great flexibility if targeting high throughput applications.

- Participants: Silviu-Ioan Filip, Matei Istoan
- Partners: Université de Rennes 1, Imperial College London
- Contact: Silviu-Ioan Filip
- URL: https://github.com/sfilip/emethod

5.4.3. *Firopt: a tool for the simultaneous design of digital FIR filters along with the dedicated hardware model*

KEYWORDS: FIR filter design, multiplierless hardware implementation generator

SCIENTIFIC DESCRIPTION: the firopt tool is an open source C++ prototype that produces Finite Impulse Response (FIR) filters that have minimal cost in terms of digital adders needed to implement them. This project aims at fusing the filter design problem from a frequency domain specification with the design of the dedicated hardware architecture. The optimality of the results is ensured by solving appropriate mixed integer linear programming (MILP) models developed for the project. It produces results that are generally more efficient than those of other methods found in the literature or from commercial tools (such as MATLAB).

- Participants: Silviu-Ioan Filip, Martin Kumm, Anastasia Volkova
- Partners: Université de Rennes 1, Université de Nantes, Fulda University of Applied Sciences
- Contact: Silviu-Ioan Filip
- URL: https://gitlab.com/filteropt/firopt

5.4.4. *Hybrid-DBT*

KEYWORDS: Dynamic Binary Translation, hardware acceleration, VLIW processor, RISC-V

SCIENTIFIC DESCRIPTION: Hybrid-DBT is a hardware/software Dynamic Binary Translation (DBT) framework capable of translating RISC-V binaries into VLIW binaries. Since the DBT overhead has to be as small as possible, our implementation takes advantage of hardware acceleration for performance critical stages (binary translation, dependency analysis and instruction scheduling) of the flow. Thanks to hardware acceleration, our implementation is two orders of magnitude faster than a pure software implementation and enable an overall performance improvements by 23% on average, compared to a native RISC-V execution.

- Participants: Simon Rokicki, Steven Derrien
- Partners: Université de Rennes 1
- URL: https://github.com/srokicki/HybridDBT

5.4.5. *Comet*

KEYWORDS: Processor core, RISC-V instruction-set architecture

SCIENTIFIC DESCRIPTION: Comet is a RISC-V pipelined processor with data/instruction caches, fully developed using High-Level Synthesis. The behavior of the core is defined in a small C code which is then fed into a HLS tool to generate the RTL representation. Thanks to this design flow, the C description can be used as a fast and cycle-accurate simulator, which behaves exactly like the final hardware. Moreover, modifications in the core can be done easily at the C level.

- Participants: Simon Rokicki, Steven Derrien, Olivier Sentieys, Davide Pala, Joseph Paturel
- Partners: Université de Rennes 1
- URL: https://gitlab.inria.fr/srokicki/Comet

5.4.6. *TypEx*

KEYWORDS: Embedded systems, Fixed-point arithmetic, Floating-point, Low power consumption, Energy efficiency, FPGA, ASIC, Accuracy optimization, Automatic floating-point to fixed-point conversion

SCIENTIFIC DESCRIPTION: TypEx is a tool designed to automatically determine custom number representations and word-lengths (i.e., bit-width) for FPGAs and ASIC designs at the C source level. The main goal of TypEx is to explore the design space spanned by possible number formats in the context of High-Level Synthesis. TypEx takes a C code written using floating-point datatypes specifying the application to be explored. The tool also takes as inputs a cost model as well as some user constraints and generates a C code where the floating-point datatypes are replaced by the wordlengths found after exploration. The best set of word-lengths is the one found by the tool that respects the given accuracy constraint and that minimizes a parametrized cost function. Figure 3 presents an overview of the TypEx design flow.

- Participants: Olivier Sentieys, Tomofumi Yuki, Van-Phu Ha
- Partners: Université de Rennes 1
- URL: https://gitlab.inria.fr/gecos/gecos-float2fix

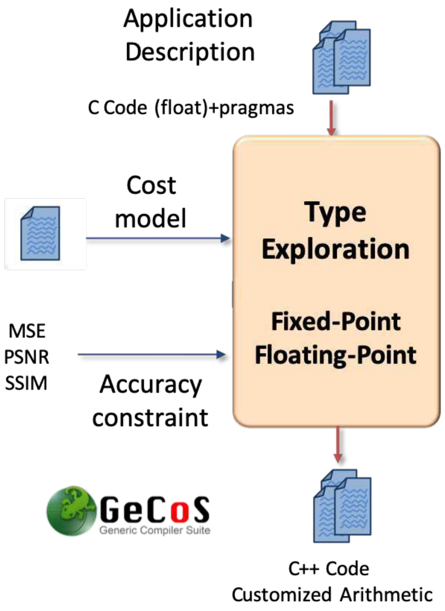

Figure 3. TypEx: a tool for type exploration and automatic floating-point to fixed-point conversion

6. New Results

6.1. Reconfigurable Architecture and Hardware Accelerator Design

6.1.1. Algorithmic Fault Tolerance for Timing Speculative Hardware

Participants: Thibaut Marty, Tomofumi Yuki, Steven Derrien.

We have been working on timing speculation, also known as overclocking, to increase the computational throughput of accelerators. However, aggressive overclocking introduces timing errors, which may corrupt the outputs to unacceptable levels. It is extremely challenging to ensure that no timing errors occur, since the probability of such errors happening depends on many factors including the temperature and process variation. Thus, aggressive timing speculation must be coupled with a mechanism to verify that the outputs are correctly computed. Our previous result demonstrated that the use of inexpensive checks based on algebraic properties of the computation can drastically reduce the cost of verifying that overclocking did not produce incorrect outputs. This has allowed the accelerator to significantly boost its throughput with little area overhead.

One weakness coming from the use of algebraic properties is that the inexpensive check is not strictly compatible with floating-point arithmetic that is not associative. This was not an issue with our previous work that targeted convolutional neural networks, which typically use fixed-point (integer) arithmetic. Our on-going work aims to extend our approach to floating-point arithmetic by using extended precision to store intermediate results, known as Kulisch accumulators. At first glance, use of extended precision that covers the full exponent range of floating-point may look costly. However, the design space of FPGAs is complex with many different trade-offs, making the optimal design highly context dependent. Our preliminary results indicate that the use of extended precision may not be any more costly than implementing the computation in floating point.

6.1.2. Adaptive Dynamic Compilation for Low-Power Embedded Systems

Participants: Steven Derrien, Simon Rokicki.

Previous works on Hybrid-DBT have demonstrated that using Dynamic Binary Translation, combined with low-power in-order architecture, enables an energy-efficient execution of compute-intensive kernels. In [32], we address one of the main performance limitations of Hybrid-DBT: the lack of speculative execution. We study how it is possible to use memory dependency speculation during the DBT process. Our approach enables fine-grained speculation optimizations thanks to a combination of hardware and software mechanisms. Our results show that our approach leads to a geo-mean speed-up of 10% at the price of a 7% area overhead. In [48], we summarize the current state of the Hybrid-DBT project and display our last results about the performance and the energy efficiency of the system. The experimental results presented here show that, for compute-intensive benchmarks, Hybrid-DBT can deliver the same performance level than a 3-issue OoO core, while consuming three times less energy. Finally, in [33], we investigate security issues caused by the use of speculation in DBT-based systems. We demonstrate that, even if those systems use in-order micro-architectures, the DBT layer optimizes binaries and speculates on the outcome of some branches, leading to security issues similar to the Spectre vulnerability. We demonstrate that both the NVidia Denver architecture and the Hybrid-DBT platform are subject to such vulnerability. However, we also demonstrate that those systems can easily be patched, as the DBT is done in software and has fine-grained control over the optimization process.

6.1.3. What You Simulate Is What You Synthesize: Designing a Processor Core from C++ Specifications

Participants: Simon Rokicki, Davide Pala, Joseph Paturel, Olivier Sentieys.

Designing the hardware of a processor core as well as its verification flow from a single high-level specification would provide great advantages in terms of productivity and maintainability. In [31] (a preliminary version also in [41]), we highlight the gain of starting from a unique high-level synthesis and simulation C++ model to design a processor core implementing the RISC-V Instruction Set Architecture (ISA). The specification code is used to generate both the hardware target design through High-Level Synthesis as well as a fast and cycle-accurate bit-accurate simulator of the latter through software compilation. The object oriented nature of C++ greatly improves the readability and flexibility of the design description compared to classical HDL-based implementations. Therefore, the processor model can easily be modified, expanded and verified using standard software development methodologies. The main challenge is to deal with C++ based synthesizable specifications of core and uncore components, cache memory hierarchy, and synchronization. In particular, the research question is how to specify such parallel computing pipelines with high-level synthesis technology and to demonstrate that there is a potential high gain in design time without jeopardizing performance and cost. Our experiments demonstrate that the core frequency and area of the generated hardware are comparable to existing RTL implementations.

6.1.4. Accelerating Itemset Sampling on FPGA

Participants: Mael Gueguen, Olivier Sentieys.

Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as ”pattern explosion”, comes from the large combinatorial space explored and is the result of too many patterns outputted to be analyzed. Recent techniques called output space sampling solve this problem by outputting only a sampled set of all the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [23], we propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. We show that our accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product. This work is done in collaboration with A. Termier from the Lacodam team at Inria.

6.1.5. Hardware Accelerated Simulation of Heterogeneous Platforms

Participants: Minh Thanh Cong, François Charot, Steven Derrien.

When considering designing heterogeneous multicore platforms, the number of possible design combinations leads to a huge design space, with subtle trade-offs and design interactions. To reason about what design is best for a given target application requires detailed simulation of many different possible solutions. Simulation frameworks exist (such as gem5) and are commonly used to carry out these simulations. Unfortunately, these are purely software-based approaches and they do not allow a real exploration of the design space. Moreover, they do not really support highly heterogeneous multicore architectures. These limitations motivate the use of hardware to accelerate the simulation of heterogeneous multicore, and in particular of FPGA components. We study an approach for designing such systems based on performance models through combining accelerator and processor core models. These models are implemented in the HAsim/LEAP infrastructure. In [21], we propose a methodology for building performance models of accelerators and describe the defined design flow.

6.1.6. Fault-Tolerant Scheduling onto Multicore embedded Systems

Participants: Emmanuel Casseau, Minyu Cui, Petr Dobias, Lei Mo, Angeliki Kritikakou.

Demand on multiprocessor systems for high performance and low energy consumption still increases in order to satisfy our requirements to perform more and more complex computations. Moreover, the transistor size gets smaller and their operating voltage is lower, which goes hand in glove with higher susceptibility to system failure. In order to ensure system functionality, it is necessary to conceive fault-tolerant systems. Temporal and/or spatial redundancy is currently used to tackle this issue. Actually, multiprocessor platforms can be less vulnerable when one processor is faulty because other processors can take over its scheduled tasks. In this context, we investigate how to map and schedule tasks onto homogeneous faulty processors.

We consider two approaches. The first approach deals with task mapping onto processors at compile time. Our goal is to guarantee both reliability and hard real-time constraints with low-energy consumption. Task duplication is assessed and duplication is performed if expected reliability of a task is not met. This work concurrently decides duplication of tasks, the task execution frequency and task allocation to minimize the energy consumption of a multicore platform with Dynamic Voltage and Frequency Scaling (DVFS) capabilities. The problem is initially formulated as Integer Non-Linear Programming and equivalently transformed to a Mixed Integer Linear Programming problem to be optimally solved. The proposed approach provides a good trade-off between energy consumption and reliability. The second approach deals with mapping and scheduling tasks at runtime. The application context is CubeSats. CubeSats operate in harsh space environment and they are exposed to charged particles and radiations, which cause transient faults. To make CubeSats fault tolerant, we propose to take advantage of their multicore architecture. We propose two online algorithms, which schedule all tasks on board of a CubeSat, detect faults and take appropriate measures (based on task replication) in order to deliver correct results. The first algorithm considers all tasks as aperiodic tasks and the second one treats them as aperiodic or periodic tasks. Their performances vary, particularly when the number of processors is low, and a choice is subject to a trade-off between the rejection rate and the energy consumption. This work is done in collaboration with Oliver Sinnen, PARC Lab., the University of Auckland.

6.1.7. Run-Time Management on Multicore Platforms

Participant: Angeliki Kritikakou.

In time-critical systems, run-time adaptation is required to improve the performance of time-triggered execution, derived based on Worst-Case Execution Time (WCET) of tasks. By improving performance, the systems can provide higher Quality-of-Service, in safety-critical systems, or execute other best-effort applications, in mixed-critical systems. To achieve this goal, we propose a parallel interference-sensitive run-time adaptation mechanism that enables a fine-grained synchronisation among cores [36]. Since the run-time adaptation of offline solutions can potentially violate the timing guarantees, we present the Response-Time Analysis (RTA) of the proposed mechanism showing that the system execution is free of timing-anomalies. The RTA takes into account the timing behavior of the proposed mechanism and its associated WCET. To support our contribution, we evaluate the behavior and the scalability of the proposed approach for different application types and execution configurations on the 8-core Texas Instruments TMS320C6678 platform. The obtained results show significant performance improvement compared to state-of-the-art centralized approaches.

6.1.8. Energy Constrained and Real-Time Scheduling and Assignment on Multicores

Participants: Olivier Sentieys, Angeliki Kritikakou, Lei Mo.

Asymmetric Multicore Processors (AMP) are a very promising architecture to deal efficiently with the wide diversity of applications. In real-time application domains, in-time approximated results are preferred to accurate – but too late – results. In [27], we propose a deployment approach that exploits the heterogeneity provided by AMP architectures and the approximation tolerance provided by the applications, so as to increase as much as possible the quality of the results under given energy and timing constraints. Initially, an optimal approach is proposed based on the problem linearization and decomposition. Then, a heuristic approach is developed based on iteration relaxation of the optimal version. The obtained results show 16.3% reduction in the computation time for the optimal approach compared to conventional optimal approaches. The proposed heuristic approach is about 100 times faster at the cost of a 29.8% QoS degradation in comparison with the optimal solution.

6.1.9. Real-Time Energy-Constrained Scheduling in Wireless Sensor and Actuator Networks

Participants: Angeliki Kritikakou, Lei Mo.

Cyber-Physical Systems (CPS), as a particular case of distributed systems, raise new challenges, because of the heterogeneity and other properties traditionally associated with Wireless Sensor and Actuator Networks (WSAN), including shared sensing, acting and real-time computing. In CPS, mobile actuators can enhance system's flexibility and scalability, but at the same time incur complex couplings in the scheduling and controlling of the actuators. In [18], we propose a novel event-driven method aiming at satisfying a required level of control accuracy and saving energy consumption of the actuators, while guaranteeing a bounded action delay. We formulate a joint-design problem of both actuator scheduling and output control. To solve this problem, we propose a two-step optimization method. In the first step, the problem of actuator scheduling and action time allocation is decomposed into two subproblems. They are solved iteratively by utilizing the solution of one in the other. The convergence of this iterative algorithm is proved. In the second step, an on-line method is proposed to estimate the error and adjust the outputs of the actuators accordingly. Through simulations and experiments,we demonstrate the effectiveness of the proposed method. In addition, many of the real-time tasks of CPS can be executed in an imprecise way. Such systems accept an approximate result as long as the baseline Quality-of-Service (QoS) is satisfied and they can execute more computations to yield better results, if more system resources are available. These systems are typically considered under the Imprecise Computation (IC) model, achieving a better tradeoff between QoS and limited system resources. However, determining a

QoS-aware mapping of these real-time IC-tasks onto the nodes of a CPS creates a set of interesting problems. In [17], we firstly propose a mathematical model to capture the dependency, energy and real-time constraints of IC-tasks, as well as the sensing, acting, and routing in the CPS. The problem is formulated as a Mixed-Integer Non-Linear Programming (MINLP) due to the complex nature of the problem. Secondly, to efficiently solve this problem, we provide a linearization method that results in a Mixed-Integer Linear Programming (MILP) formulation of our original problem. Finally, we decompose the transformed problem into a task allocation subproblem and a task adjustment subproblem, and, then, we find the optimal solution based on subproblem iteration. Through the simulations, we demonstrate the effectiveness of the proposed method. Last, but not least, wireless charging can provide dynamic power supply for CPS. Such systems are typically considered under the scenario of Wireless Rechargeable Sensor Networks (WRSNs). With the use of Mobile Chargers (MCs), the flexibility of WRSNs is further enhanced. However, the use of MCs poses several challenges during the system design. The coordination process has to simultaneously optimize the scheduling, the moving time and the charging time of multiple MCs, under limited system resources (e.g., time and energy). Efficient methods that jointly solve these challenges are generally lacking in the literature. In [16], we address the multiple MCs coordination problem under multiple system requirements. Firstly, we aim at minimizing the energy consumption of MCs, guaranteeing that every sensor will not run out of energy. We formulate the multiple MCs coordination problem as a mixed-integer linear programming and derive a set of desired network properties. Secondly, we propose a novel decomposition method to optimally solve the problem, as well as to reduce the computation time. Our approach divides the problem into a subproblem for the MC scheduling and a subproblem for the MC moving time and charging time, and solves them iteratively by utilizing the solution of one into the other. The convergence of the proposed method is analyzed theoretically. Simulation results demonstrate the effectiveness and scalability of the proposed method in terms of solution quality and computation time.

6.1.10. Fault-Tolerant Microarchitectures

Participants: Joseph Paturel, Angeliki Kritikakou, Olivier Sentieys.

As transistors scale down, processors are more vulnerable to radiation that can cause multiple transient faults in function units. Rather than excluding these units from execution, performance overhead of VLIW processors can be reduced when fault-free components of these affected units are still used. In [29], the function units are enhanced with coarse-grained fault detectors. A re-scheduling of the instructions is performed at run-time to use not only the healthy function units, but also the fault-free components of the faulty function units. The scheduling window of the proposed mechanism covers two instruction bundles, which makes it suitable to explore mitigation solutions in the current and in the next instruction execution. Experiments show that the proposed approach can mitigate a large number of faults with low performance and area overheads. In addition, technology scaling can cause transient faults with long duration. In this case, the affected function unit is usually considered as faulty and is not further used. To reduce this performance degradation, we proposed a hardware mechanism to (i) detect the faults that are still active during execution and (ii) re-schedule the instructions to use the fault-free components of the affected function units [30]. When the fault faints, the affected function unit components can be reused. The scheduling window of the proposed mechanism is two instruction bundles being able to exploit function units of both the current and the next instruction execution. The results show multiple long-duration fault mitigation can be achieved with low performance, area, and power overhead.

Simulation-based fault injection is commonly used to estimate system vulnerability. Existing approaches either partially model the studied system's fault masking capabilities, losing accuracy, or require prohibitive estimation times. Our work proposes a vulnerability analysis approach that combines gate-level fault injection with microarchitecture-level Cycle-Accurate and Bit-Accurate simulation, achieving low estimation time. Faults both in sequential and combinational logic are considered and fault masking is modeled at gate-level, microarchitecture-level and application-level, maintaining accuracy. Our case-study is a RISC-V processor. Obtained results show a more than 8% reduction in masked errors, increasing more than 55% system failures compared to standard fault injection approaches. This work is currently under review.

6.1.11. Fault-Tolerant Networks-on-Chip

Participants: Romain Mercier, Cédric Killian, Angeliki Kritikakou, Daniel Chillet.

Network-on-Chip has become the main interconnect in the multicore/manycore era since the beginning of this decade. However, these systems become more sensitive to faults due to transistor shrinking size. In parallel, approximate computing appears as a new computation model for applications since several years. The main characteristic of these applications is to support the approximation of data, both for computations and for communications. To exploit this specific application property, we develop a fault-tolerant NoC to reduce the impact of faults on the data communications. To address this problem, we consider multiple permanent faults on router which cannot be managed by Error-Correcting Codes (ECCs) and we propose a bit-shuffling method to reduce the impact of faults on Most Significant Bits (MSBs), hence permanent faults only impact Low Significant Bits (LSBs) instead of MSBs reducing the errors impact. We evaluated the proposed method for data mining benchmark and we show that our proposal can lead to 73.04% reduction on the clustering error rate and 84.64% reduction on the mean centroid Mean Square Error (MSE) for 3-bit permanent faults which affect MSBs on 32-bit words with a limited area cost. This work is currently under review for an international conference.

6.1.12. Improving the Reliability of Wireless Network-on-Chip (WiNoC)

Participants: Joel Ortiz Sosa, Olivier Sentieys, Cédric Killian.

Wireless Network-on-Chip (WiNoC) is one of the most promising solutions to overcome multi-hop latency and high power consumption of modern many/multi core System-on-Chip (SoC). However, standard WiNoC approaches are vulnerable to multi-path interference introduced by on-chip physical structures. To overcome such parasitic phenomenon, we first proposed a Time-Diversity Scheme (TDS) to enhance the reliability of on-chip wireless links using a realistic wireless channel model. We then proposed an adaptive digital transceiver, which enhances communication reliability under different wireless channel configurations in [38]. Based on the same realistic channel model, we investigated the impact of using some channel correction techniques. Experimental results show that our approach significantly improves Bit Error Rate (BER) under different wireless channel configurations. Moreover, our transceiver is designed to be adaptive, which allows for wireless communication links to be established in conditions where this would not be possible for standard transceiver architectures. The proposed architecture, designed using a 28-nm FDSOI technology, consumes only 3.27 mW for a data rate of 10 Gbit/s and has a very small area footprint. We also proposed a low-power, high-speed, multi-carrier reconfigurable transceiver based on Frequency Division Multiplexing (FDM) to ensure data transfer in future Wireless NoCs in [37]. The proposed transceiver supports a medium access control method to sustain unicast, broadcast and multicast communication patterns, providing dynamic data exchange among wireless nodes. Designed using a 28-nm FDSOI technology, the transceiver only consumes 2.37 mW and 4.82 mW in unicast/broadcast and multicast modes, respectively, with an area footprint of 0.0138 mm^2.

6.1.13. Error Mitigation in Nanophotonic Interconnect

Participants: Jaechul Lee, Cédric Killian, Daniel Chillet.

The energy consumption of manycore is dominated by data movements, which calls for energy-efficient and high-bandwidth interconnects. Integrated optics is promising technology to overcome the bandwidth limitations of electrical interconnects. However, it suffers from high power overhead related to low efficiency lasers, which calls for the use of approximate communications for error tolerant applications. In this context, in [25] we investigate the design of an Optical NoC supporting the transmission of approximate data. For this purpose, the least significant bits of floating point numbers are transmitted with low power optical signals. A transmission model allows estimating the laser power according to the targeted BER and a micro-architecture allows configuring, at run-time, the number of approximated bits and the laser output powers. Simulation results show that, compared to an interconnect involving only robust communications, approximations in the optical transmissions lead to a laser power reduction up to 42% for image processing application with a limited degradation at the application level.

6.2. Compilation and Synthesis for Reconfigurable Platform

6.2.1. Compile Time Simplification of Sparse Matrix Code Dependences

Participant: Tomofumi Yuki.

In [28], we developed a combined compile-time and runtime loop-carried dependence analysis of sparse matrix codes and evaluated its performance in the context of wavefront parallellism. Sparse computations incorporate indirect memory accesses such as x[col[j]] whose memory locations cannot be determined until runtime. The key contributions are two compile-time techniques for significantly reducing the overhead of runtime dependence testing: (1) identifying new equality constraints that result in more efficient runtime inspectors, and (2) identifying subset relations between dependence constraints such that one dependence test subsumes another one that is therefore eliminated. New equality constraints discovery is enabled by taking advantage of domain-specific knowledge about index arrays, such as col[j]. These simplifications lead to automatically-generated inspectors that make it practical to parallelize such computations. We analyze our simplification methods for a collection of seven sparse computations. The evaluation shows our methods reduce the complexity of the runtime inspectors significantly. Experimental results for a collection of five large matrices show parallel speedups ranging from 2x to more than 8x running on a 8-core CPU.

6.2.2. Study of Polynomial Scheduling

Participant: Tomofumi Yuki.

We have studied the Handelman's theorem used for polynomial scheduling, which resembles the Farkas' lemma for affine scheduling. Theorems from real algebraic geometry and polynomial optimization show that some polynomials have Handelman representations when they are non-negative on a domain, instead of strictly positive as stated in Handelman's theorem. The global minimizers of a polynomial must be at the boundaries of the domain to have such a representation with finite bounds on the degree of monomials. This creates discrepancies in terms of polynomials included in the exploration space with a fixed bound on the monomial degree. Our findings give an explanation to our failed attempt to apply polynomial scheduling to Index-Set Splitting: we were precisely trying to find polynomials with global minimizers at the interior of a domain.

6.2.3. Optimizing and Parallelizing compilers for Time-Critical Systems

Participant: Steven Derrien.

6.2.3.1. Contentions-Aware Task-Level Parallelization

Accurate WCET analysis for multicores is challenging due to concurrent accesses to shared resources, such as communication through bus or Network on Chip (NoC). Current WCET techniques either produce pessimistic WCET estimates or preclude conflicts by constraining the execution, at the price of a significant hardware under-utilization. Most existing techniques are also restricted to independent tasks, whereas real-time workloads will probably evolve toward parallel programs. The WCET behavior of such parallel programs is even more challenging to analyze because they consist of *dependent* tasks interacting through complex synchronization/communication mechanisms. In [35], we propose a scheduling technique that jointly selects Scratchpad Memory (SPM) contents off-line, in such a way that the cost of SPM loading/unloading is hidden. Communications are fragmented to augment hiding possibilities. Experimental results show the effectiveness of the proposed technique on streaming applications and synthetic task-graphs. The overlapping of communications with computations allows the length of generated schedules to be reduced by 4% on average on streaming applications, with a maximum of 16%, and by 8% on average for synthetic task graphs. We further show on a case study that generated schedules can be implemented with low overhead on a predictable multicore architecture (Kalray MPPA).

6.2.3.2. WCET-Aware Parallelization of Model-Based Applications for Multicores

Parallel architectures are nowadays increasingly used in embedded time-critical systems. The Argo H2020 project provides a programming paradigm and associated tool flow to exploit the full potential of architectures in terms of development productivity, time-to-market, exploitation of the platform computing power and guaranteed real-time performance. The Argo toolchain operates on Scilab and XCoS inputs, and targets ScratchPad Memory (SPM)-based multicores. Data-layout and loop transformations play a key role in this flow as they improve SPM efficiency and reduce the number of accesses to shared main memory. In [19] we present the overall results of the project, a compiler tool-flow for automated parallelization of model-based real-time software, which addresses the shortcomings of multi-core architectures in real-time systems. The flow is demonstrated using a model-based Terrain Awareness and Warning Systems (TAWS) and an edge detection algorithm from the image-processing domain. Model-based applications are first transformed into real-time C code and from there into a well-predictable parallel C program. Tight bounds for the Worst-Case Execution Time (WCET) of the parallelized program can be determined using an integrated multicore WCET analysis. Thanks to the use of an architecture description language, the general approach is applicable to a wider range of target platforms. An experimental evaluation for a research architecture with network-on-chip (NoC) interconnect shows that the parallel WCET of the TAWS application can be improved by factor 1.77 using the presented compiler tools.

6.2.3.3. WCET oriented Iterative compilation

Static Worst-Case Execution Time (WCET) estimation techniques operate upon the binary code of a program in order to provide the necessary input for schedulability analysis techniques. Compilers used to generate this binary code include tens of optimizations, that can radically change the flow information of the program. Such information is hard to maintain across optimization passes and may render automatic extraction of important flow information, such as loop bounds, impossible. Thus, compiler optimizations, especially the sophisticated optimizations of mainstream compilers, are typically avoided. In this work, published in [22], we explore for the first time iterative-compilation techniques that reconcile compiler optimizations and static WCET estimation. We propose a novel learning technique that selects sequences of optimizations that minimize the WCET estimate of a given program. We experimentally evaluate the proposed technique using an industrial WCET estimation tool (AbsInt aiT) over a set of 46 benchmarks from four different benchmarks suites, including reference WCET benchmark applications, image processing kernels and telecommunication applications. Experimental results show that WCET estimates are reduced on average by 20.3% using the proposed technique,as compared to the best compiler optimization level applicable.

6.2.4. Towards Generic and Scalable Word-Length Optimization

Participants: Van-Phu Ha, Tomofumi Yuki, Olivier Sentieys.

Fixed-Point arithmetic is widely used for implementing Digital Signal Processing (DSP) systems on electronic devices. Since initial specifications are often written using floating-point arithmetic, conversion to fixed-point is a recurring step in hardware design. The primary objective of this conversion is to minimize the cost (energy and/or area) while maintaining an acceptable level of quality at the output. In Word-Length Optimization (WLO), each variable/operator may be assigned a different fixed-point encoding, which means that the design space grows exponentially as the number of variables increases. This is especially true when targeting hardware accelerators implemented in FPGA or ASIC. Thus, most approaches for WLO involve heuristic search algorithms. In [24] (a preliminary version also in [40]), we propose a method to improve the scalability of Word-Length Optimization (WLO) for large applications that use complex quality metrics such as Structural Similarity (SSIM). The input application is decomposed into smaller kernels to avoid uncontrolled explosion of the exploration time, which is known as noise budgeting. The main challenge addressed in this paper is how to allocate noise budgets to each kernel. This requires capturing the interactions across kernels. The main idea is to characterize the impact of approximating each kernel on accuracy/cost through simulation and regression. Our approach improves the scalability while finding better solutions for Image Signal Processor pipeline.

In [26], we propose an analytical approach to study the impact of floating-point (FlP) precision variation on the square root operation, in terms of computational accuracy and performance gain. We estimate the round-off error resulting from reduced precision. We also inspect the Newton Raphson algorithm used to approximate the square root in order to bound the error caused by algorithmic deviation. Consequently, the implementation of the square root can be optimized by fittingly adjusting its number of iterations with respect to any given FlP precision specification, without the need for long simulation times. We evaluate our error analysis of the square root operation as part of approximating a classic data clustering algorithm known as K-means, for the purpose of reducing its energy footprint. We compare the resulting inexact K-means to its exact counterpart, in the context of color quantization, in terms of energy gain and quality of the output. The experimental results show that energy savings could be achieved without penalizing the quality of the output (e.g., up to 41.87% of energy gain for an output quality, measured using structural similarity, within a range of [0.95,1]).

6.2.5. *Optimized Implementations of Constant Multipliers for FPGAs*

Participant: Silviu-Ioan Filip.

The multiplication by a constant is a frequently used arithmetic operation. To implement it on Field Programmable Gate Arrays(FPGAs), the state of the art offers two completely different methods: one relying on bit shifts and additions/subtractions, and another one using look-up tables and additions. So far, it was unclear which method performs best for a given constant and input/output data types. The main contribution of the work published in [39] is a thorough comparison of both methods in the main application contexts of constant multiplication: filters, signal-processing transforms, and elementary functions. Most of the previous state of the art addresses multiplication by an integer constant. This work shows that, in most of these application contexts, a formulation of the problem as the multiplication by a real constant allows for more efficient architectures. Another contribution is a novel extension of the shift-and-add method to real constants. For that, an integer linear programming (ILP) formulation is proposed, which truncates each component in the shift-and-add network to a minimum necessary word size that is aligned with the approximation error of the coefficient. All methods are implemented within the open-source FloPoCo framework.

6.2.6. *Optimal Multiplierless FIR Filter Design*

Participant: Silviu-Ioan Filip.

The hardware optimization of direct form finite impulse response (FIR) filters has been a topic of research for the better part of the last four decades and is still garnering significant research and industry interest. In [47], we present two novel optimization methods based on integer linear programming (ILP) that minimize the number of adders used to implement a direct/transposed FIR filter adhering to a given frequency specification. The proposed algorithms work by either fixing the number of adders used to implement the products (multiplier block adders) or by bounding the adder depth (AD) used for these products. The latter can be used to design filters with minimal AD for low power applications. In contrast to previous multiplierless FIR approaches, the methods introduced here ensure adder count optimality. To demonstrate their effectiveness, we perform several experiments using established design problems from the literature, showing superior results.

6.2.7. *Application-specific arithmetic in high-level synthesis tools*

Participant: Steven Derrien.

In [49], we have shown that the use of non-conventional implementation for floating-point arithmetic can bring significant benefits when used in the context of High-Level Synthesis. We are currently building on these preliminary results to show that it is possible to implement accelerators using exact floating-point arithmetic for similar performance/area cost than standard floating-point operators implementations. Our approach builds on Kulish's approach to implement floating-point adders, and targets dense Matrix Products kernels (GEM3 like) accelerators on FPGAs.

6.3. Applications

6.3.1. *SmartSense*

Participants: Nicolas Roux, Olivier Sentieys.

Developing smarter and greener buildings has been an expanding field of research over the last decades. One of the essential requirements for energy utilities is the knowledge of power consumption patterns at the single-appliance level. To estimate these patterns without using an individual power meter for each appliance, Non-Intrusive Load Monitoring (NILM) consists in disaggregating electrical loads by examining the appliance specific power consumption signature within the aggregated load single measurement. Therefore, the method is considered non-intrusive since the data are collected from a single electrical panel outside of the monitored building. Thus, NILM has been a very active field of research with renewed interest over the last years.

Therefore, knowing the plug-level power consumption of each appliance in a building can lead to drastic savings in energy consumption. In [34], we have addressed the issue of NILM inaccuracy in the context of industrial or commercial buildings, by combining data from a low-cost, general-purpose, wireless sensor network. We have proposed a novel approach based on a simplex solver to estimate the power load values of the steady states on sliding windows of data with varying size. We have shown the principle of the approach and demonstrated its interest, limited complexity, and ease of use.

7. Partnerships and Cooperations

7.1. Regional Initiatives

7.1.1. Labex CominLabs - BBC (2016-2020)

Participants: Olivier Sentieys, Cédric Killian, Joel Ortiz Sosa.

The aim of the BBC (on-chip wireless Broadcast-Based parallel Computing) project is to evaluate the use of wireless links between cores inside chips and to define new paradigms. Using wireless communications enables broadcast capabilities for Wireless Networks on Chip (WiNoC) and new management techniques for memory hierarchy and parallelism. The key objectives concern improvement of power consumption, estimation of achievable data rates, flexibility and reconfigurability, size reduction and memory hierarchy management. In this project, CAIRN is addressing new low-power MAC (media access control) technique based on CDMA access as well as broadcast-based fast cooperation protocol designed for resource sharing (bandwidth, distributed memory, cache coherency) and parallel programming. For more details see https://bbc.cominlabs.u-bretagneloire.fr

7.2. National Initiatives

7.2.1. ANR AdequateDL

Participants: Olivier Sentieys, Silviu-Ioan Filip.

Program: ANR PRC
Project acronym: AdequateDL
Project title: Approximating Deep Learning Accelerators
Duration: Jan. 2019 - Dec. 2022
Coordinator: Cairn
Other partners: INL, CAIRN, LIRMM, CEA-LIST

The design and implementation of convolutional neural networks for deep learning is currently receiving a lot of attention from both industrials and academics. However, the computational workload involved with CNNs is often out of reach for low power embedded devices and is still very costly when run on datacenters. By relaxing the need for fully precise operations, approximate computing substantially improves performance and energy efficiency. Deep learning is very relevant in this context, since playing with the accuracy to reach adequate computations will significantly enhance performance, while keeping quality of results in a user-constrained range. AdequateDL will explore how approximations can improve performance and energy efficiency of hardware accelerators in deep-learning applications. Outcomes include a framework for accuracy exploration and the demonstration of order-of-magnitude gains in performance and energy efficiency of the proposed adequate accelerators with regards to conventional CPU/GPU computing platforms.

7.2.2. ANR RAKES

Participants: Olivier Sentieys, Cédric Killian, Joel Ortiz Sosa.

Program: ANR PRC

Project acronym: RAKES

Project title: Radio Killed an Electronic Star: speed-up parallel programming with broadcast communications based on hybrid wireless/wired network on chip

Duration: June 2019 - June 2023

Coordinator: TIMA

Other partners: TIMA, CAIRN, Lab-STICC

The efficient exploitation by software developers of multi/many-core architectures is tricky, especially when the specificities of the machine are visible to the application software. To limit the dependencies to the architecture, the generally accepted vision of the parallelism assumes a coherent shared memory and a few, either point to point or collective, synchronization primitives. However, because of the difference of speed between the processors and the main memory, fast and small dedicated hardware controlled memories containing copies of parts of the main memory (a.k.a caches) are used. Keeping these distributed copies up-to-date and synchronize the accesses to shared data, requires to distribute and share information between some may if not all the nodes. By nature, radio communications provide broadcast capabilities at negligible latency, they have thus the potential to disseminate information very quickly at the scale of a circuit and thus to be an opening for solving these issues. In the RAKES project, we intend to study how wireless communications can solve the scalability of the abovementioned problems, by using mixed wired/wireless Network on Chip. We plan to study several alternatives and to provide (a) a virtual platform for evaluation of the solutions and (b) an actual implementation of the solutions.

7.2.3. ANR Opticall^2

Participants: Olivier Sentieys, Cédric Killian, Daniel Chillet.

Program: ANR PRCE

Project acronym: Opticall^2

Project title: on-chip OPTIcal interconnect for ALL to ALL communications

Duration: Dec. 2018 - Nov. 2022

Coordinator: INL

Other partners: INL, CAIRN, C2N, CEA-LETI, Kalray

The aim of Opticall^2 is to design broadcast-enabled optical communication links in manycore architectures at wavelengths around 1.3μm. We aim to fabricate an optical broadcast link for which the optical power is equally shared by all the destinations using design techniques (different diode absorption lengths, trade-off depending on the current point in the circuit and the insertion losses). No optical switches will be used, which will allow the link latency to be minimized and will lead to deterministic communication times, which are both key features for efficient cache coherence protocols. The second main objective of Opticall^2 is to propose and design a new broadcast-aware cache coherence communication protocol allowing hundreds of computing clusters and memories to be interconnected, which is well adapted to the broadcast-enabled optical communication links. We expect better performance for the parallel execution of benchmark programs, and lower overall power consumption, specifically that due to invalidation or update messages.

7.2.4. ANR SHNOC

Participants: Cédric Killian, Daniel Chillet, Olivier Sentieys, Emmanuel Casseau.

Program: ANR JCJC (young researcher)

Project acronym: SHNOC

Project title: Scalable Hybrid Network-on-Chip

Duration: Feb. 2019 - Jan. 2022

P.I.: C. Killian, CAIRN

The goal of the SHNoC project is to tackle one of the manycore interconnect issues (scalability in terms of energy consumption and latency provided by the communication medium) by mixing emerging technologies. Technology evolution has allowed for the integration of silicon photonics and wireless on-chip communications, creating Optical and Wireless NoCs (ONoCs and WNoCs, respectively) paradigms. The recent publications highlight advantages and drawbacks for each technology: WNoCs are efficient for broadcast, ONoCs have low latency and high integrated density (throughput/cm^2) but are inefficient in multicast, while ENoCs are still the most efficient solution for small/average NoC size. The first contribution of this project is to study the compatibility of processes to associate the three aforementioned technologies and to define an hybrid topology of the interconnection architecture. This exploration will determine the number of antennas for the WNoC, the amount of embedded lasers sources for the ONoC and the routers architecture for the ENoC. The second main contribution is to provide quality of service of communication by determining, at run-time, the best path among the three NoCs with respect to a target, e.g. minimizing the latency or energy. We expect to demonstrate that the three technologies are more efficient when jointly used and combined, with respect to traffic characteristics between cores and quality of service targeted.

7.2.5. IPL ZEP

Participants: Davide Pala, Olivier Sentieys.

Program: Inria Project Lab

Project acronym: ZEP

Project title: Zero-Power Computing Systems

Duration: Oct. 2017 - Nov. 2020

Coordinator: Inria Socrate

Other partners: Pacap, Cairn, Corse, CEA-LETI

The ZEP project addresses the issue of designing tiny, batteryless, computing objects harvesting energy in the environment. The main application target is Internet of Things (IoT) where small communicating objects will be composed of this computing part associated to a low-power wake-up radio system. The energy level harvested being very low, very frequent energy shortages are expected, which makes the systems following the paradigm of Intermittently-Powered Systems. In order for the system to maintain a consistent state, it will be based on a new architecture embedding non-volatile memory (NVRAM). The major outcomes of the project will be a prototype harvesting board including NVRAM and the design of a new non-volatile processor (NVP) associated with its optimizing compiler and operating system. Cairn is focusing on the microarchitecture of the NVP and on new strategies for backup and restore data and processor state. The ZEP project gathers four Inria teams that have a scientific background in architecture, compilation, operating system and low power together with the CEA Grenoble. Another important goal of the project is to structure the research and innovation that should occur within Inria to prepare the important technological shift brought by NVRAM technologies.

7.2.6. DGA RAPID - FLODAM (2017–2021)

Participants: Joseph Paturel, Simon Rokicki, Olivier Sentieys, Angeliki Kritikakou.

FLODAM is an industrial research project for methodologies and tools dedicated to the hardening of embedded multi-core processor architectures. The goal is to: 1) evaluate the impact of the natural or artificial environments on the resistance of the system components to faults based on models that reflect the reality of the system environment, 2) the exploration of architecture solutions to make the multi-core architectures fault tolerant to transient or permanent faults, and 3) test and evaluate the proposed fault tolerant architecture solutions and compare the results under different scenarios provided by the fault models. For more details see https://flodam.fr

7.3. European Initiatives

7.3.1. H2020 ARGO

Participants: Steven Derrien, Angeliki Kritikakou, Olivier Sentieys.

Program: H2020-ICT-04-2015

Project acronym: ARGO

Project title: WCET-Aware Parallelization of Model-Based Applications for Heterogeneous Parallel Systems

Duration: Feb. 2016 - Feb. 2019

Coordinator: KIT

Other partners: KIT (Germany), UR1/Inria/CAIRN, Recore Systems (Netherlands), TEI-WG (Greece), Scilab Ent. (France), Absint (Ger.), DLR (Ger.), Fraunhofer (Ger.)

Increasing performance and reducing cost, while maintaining safety levels and programmability are the key demands for embedded and cyber-physical systems, e.g. aerospace, automation, and automotive. For many applications, the necessary performance with low energy consumption can only be provided by customized computing platforms based on heterogeneous many-core architectures. However, their parallel programming with time-critical embedded applications suffers from a complex toolchain and programming process. ARGO will address this challenge with a holistic approach for programming heterogeneous multi- and many-core architectures using automatic parallelization of model-based real-time applications. ARGO will enhance WCET-aware automatic parallelization by a cross-layer programming approach combining automatic tool-based and user-guided parallelization to reduce the need for expertise in programming parallel heterogeneous architectures. The ARGO approach will be assessed and demonstrated by prototyping comprehensive time-critical applications from both aerospace and industrial automation domains on customized heterogeneous many-core platforms.

7.3.2. ANR International ARTEFaCT

Participants: Olivier Sentieys, Van-Phu Ha, Tomofumi Yuki.

Program: ANR International France-Switzerland

Project acronym: ARTEFaCT

Project title: AppRoximaTivE Flexible Circuits and Computing for IoT

Duration: Feb. 2016 - Dec. 2019

Coordinator: CEA

Other partners: CEA-LETI, CAIRN, EPFL

The ARTEFaCT project aims to build on the preliminary results on inexact and exact near-threshold and sub-threshold circuit design to achieve major energy consumption reductions by enabling adaptive accuracy control of applications. ARTEFaCT proposes 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 applications. The project will contribute along three research directions: (1) approximate, ultra low-power circuit design, (2) modeling and analysis of variable levels of computation precision in applications, and (3) accuracy-energy trade- offs in software.

7.4. International Initiatives

7.4.1. Inria International Labs

EPFL-Inria

Associate Team involved in the International Lab:

7.4.1.1. IoTA

Title: Ultra-Low Power Computing Platform for IoT leveraging Controlled Approximation

International Partner (Institution - Laboratory - Researcher):

Ecole Polytechnique Fédérale de Lausanne (Switzerland) - Prof. Christian Enz

Start year: 2017

See also: https://team.inria.fr/cairn/IOTA

Energy issues are central to the evolution of the Internet of Things (IoT), and more generally to the ICT industry. Current low-power design techniques cannot support the estimated growth in number of IoT objects and at the same time keep the energy consumption within sustainable bounds, both on the IoT node side and on cloud/edge-cloud side. This project aims to build on the preliminary results on inexact and exact sub/near-threshold circuit design to achieve major energy consumption reductions by enabling adaptive accuracy control of applications. Advanced ultra low-power hardware design methods utilize very low supply voltage, such as in near-threshold and sub-threshold designs. These emerging technologies are very promising avenues to decrease active and stand-by-power in electronic devices. To move another step forward, recently, approximate computing has become a major field of research in the past few years. IoTA proposes to address, in a consistent fashion, the entire design stack, from 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. The main scientific challenge is twofold: (1) to add adaptive accuracy to hardware blocks built in near/sub threshold technology and (2) to provide the tools and methods to program and make efficient use of these hardware blocks for applications in the IoT domain. This entails developing approximate computing units, on one side, and methods and tools, on the other side, to rigorously explore trade-offs between accuracy and energy consumption in IoT systems. The expertise of the members of the two teams is complementary and covers all required technical knowledge necessary to reach our objectives, i.e., ultra low power hardware design (EPFL), approximate operators and functions (Inria, EPFL), formal analysis of precision in algorithms (Inria), and static and dynamic energy management (Inria, EPFL). Finally, the proof of concept will consist of results on (1) an adaptive, inexact or exact, ultra-low power microprocessor in 28 nm process and (2) a real prototype implemented in an FPGA platform combining processors and hardware accelerators. Several software use-cases relevant for the IoT domain will be considered, e.g., embedded vision, IoT sensors data fusion, to practically demonstrate the benefits of our approach.

7.4.2. Inria Associate Teams Not Involved in an Inria International Labs

7.4.2.1. IntelliVIS

Title: Design Automation for Intelligent Vision Hardware in Cyber Physical Systems

International Partner (Institution - Laboratory - Researcher):

IIT Goa (India) - Prof. Sharad Sinha

Start year: 2019

The proposed collaborative research work is focused on the design and development of artificial intelligence based embedded vision architectures for cyber physical systems (CPS). Embedded vision architectures for cyber physical systems (CPS), sometimes referred to as “Visual IoT”, are challenging to design because of primary constraints of compute resources, energy and power management. Embedded vision nodes in CPS, when designed with the application of Artificial Intelligence principles and algorithms, will turn into intelligent nodes (self-learning devices) capable of performing computation and inference at the node resulting in node-level cognition. This would allow only necessary and relevant post processed data to be sent to a human or a computer-based analyst for further processing and refinement in results. However, design and development of such nodes is non-trivial. Many existing computer vision algorithms, typically ported to embedded platforms, are compute and memory intensive thus limiting the operational time when ported to battery powered devices. In addition, transmission of captured visual data, with minimal processing at the node to extract actionable insights poses increased demands on computational, communication and energy requirements. Visual saliency i.e. extraction of key features or regions of interest in images or videos captured by an embedded vision node and related post processing for inference using AI techniques is an interesting and challenging research direction. The primary reason being that such an approach is expected to cover a wider range of application specific scenarios than statically determined approaches specific to each scenario involving remote off-loading of compute or scenario specific data on servers. Apart from a general approach to visual saliency in nodes using AI based methods (machine and deep learning methods), another principal goal of the proposed project is also to examine and propose methods that allow rapid deployment of AI techniques in these nodes. Many AI techniques are data driven and for a node to adapt from one environment or application specific scenario to another, rapid deployment of AI techniques over the air (OTA) would be an interesting and challenging research direction.

7.4.3. Inria International Partners

7.4.3.1. DARE

Title: Design space exploration Approaches for Reliable Embedded systems

International Partner (Institution - Laboratory - Researcher):

IMEC (Belgium) - Francky Catthoor, IMEC fellow

Duration: 2017 - 2021

Start year: 2017

This collaborative research focuses on methodologies to design low cost and efficient techniques for safety-critical embedded systems, which require high performance and safety implying both fault tolerance and hard real-time constraints. More precisely, the objective is to develop Design Space Exploration (DSE) methodology applicable to any platform domain to drive the design of adaptive predictable low cost and efficient error detection techniques. Run-time dynamic control mechanisms are proposed to actively optimize system fault tolerance by exploring the trade-offs between predictability, reliability, performance and energy consumption using the information received from the environment and the platform during execution. In contrast to design-time static approaches the dynamism can then be exploited to improve energy consumption and performance.

7.4.3.2. LRS

Title: Loop unRolling Stones: compiling in the polyhedral model

International Partner (Institution - Laboratory - Researcher):

Colorado State University (United States) - Department of Computer Science - Prof. Sanjay Rajopadhye

7.4.3.3. HARAMCOP

Title: Hardware accelerators modeling using constraint-based programming

International Partner (Institution - Laboratory - Researcher):

Lund University (Sweden) - Department of Computer Science - Prof. Krzysztof Kuchcinski

7.4.3.4. DeLeES

Title: Energy-efficient Deep Learning Systems for Low-cost Embedded Systems

International Partner (Institution - Laboratory - Researcher):

University of British Columbia (Vancouver, Canada) - Electrical and Computer Engineering - Prof. Guy Lemieux

Start year: 2018

This collaboration is centered around creation of deep-learning inference systems which are energy efficient and low cost. There are two design approaches: (i) an all-digital low-precision system, and (ii) mixed analog/digital low-precision system.

7.4.3.5. Informal International Partners

Dept. of Electrical and Computer Engineering, Concordia University (Canada), Optical network-on-chip, manycore architectures.

LSSI laboratory, Québec University in Trois-Rivières (Canada), Design of architectures for digital filters and mobile communications.

Department of Electrical and Computer Engineering, University of Patras (Greece), Wireless Sensor Networks, Worst-Case Execution Time, Priority Scheduling.

Karlsruhe Institute of Technology - KIT (Germany), Loop parallelization and compilation techniques for embedded multicores.

PARC Lab., the University of Auckland (New-Zealand), Fault-tolerant task scheduling onto multi-core.

Ruhr - University of Bochum - RUB (Germany), Reconfigurable architectures.

University of Science and Technology of Hanoi (Vietnam), Participation of several CAIRN's members in the Master ICT / Embedded Systems.

7.5. International Research Visitors

7.5.1. Visits of International Scientists

- Bernard Goossens, Univ. Perpignan, July 2019.
- Sharad Sinha, IIT Goa, India, July 2019.

7.5.2. Visits to International Teams

7.5.2.1. Sabbatical programme

Steven Derrien visited Colorado State University for a 6 month sabbatical from January to July 2019, where he collaborated with Sanjay Rajopadhye. This collaboration has led to two joint PhD between Université de Rennes 1 and Colorado State University which both started in late 2019.

7.5.2.2. Research Stays Abroad

- Olivier Sentieys visited Colorado State University, Computer Science Department and gave a seminar on Approximate Computing in November 2019.
- P. Dobias (PhD student) spent 5 months in the Parallel and Reconfigurable Lab. of the Electrical and Computer Engineering department, the University of Auckland, New Zealand, from November 2018 until March 2019.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events: Organisation

8.1.1.1. General Chair, Scientific Chair

- D. Chillet was the General Co-Chair of HiPEAC RAPIDO'19 Workshop.

8.1.1.2. Member of the Organizing Committees

- E. Casseau is a member of DASIP Steering Committee, Conference on Design and Architectures for Signal and Image Processing.

8.1.2. Scientific Events: Selection

8.1.2.1. Chair of Conference Program Committees

- O. Sentieys is Co-Chair of the D8 Track on Architectural and Microarchitectural Design at IEEE/ACM DATE since 2018.
- O. Sentieys is a member of the committee for delivering the Best Paper Award at IEEE/ACM DATE 2020.
- O. Sentieys served as a committee member in the IEEE EDAA Outstanding Dissertations Award (ODA).

8.1.2.2. Member of the Conference Program Committees

- D. Chillet was member of the technical program committee of HiPEAC RAPIDO, HiPEAC WRC, MCSoC, DCIS, ComPAS, DASIP, LP-EMS, ARC.
- S. Derrien was a member of technical program committee of IEEE FPL'19, IEEE FPT'19, IEEE ASAP'19 and ARC'19.
- A. Kritikakou was a member of technical program committee of IEEE RTAS'20, ECRTS'19, SAMOS'19, DATE'20.
- O. Sentieys was a member of technical program committee of IEEE/ACM DATE, IEEE FPL, ACM ENSSys, ACM SBCCI, IEEE ReConFig.
- T. Yuki was a member of technical program committee of CGO '19, SC '19, TAPAS '19, CC '20, IMPACT '20, and was a member of external review committee of PACT '19.

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

- D. Chillet is member of the Editor Board of Journal of Real-Time Image Processing (JRTIP).
- O. Sentieys is member of the editorial board of Journal of Low Power Electronics.

8.1.4. Invited Talks

- O. Sentieys gave an invited talk at FETCH (École d'hiver Francophone sur les Technologies de Conception des Systèmes embarqués Hétérogènes), Louvain-la-Neuve, Belgique, January 2019 on "Approximating Deep Learning Accelerators".
- O. Sentieys gave an invited talk at ARCHI Spring School, Lorient, France, May 2019 on "Design of VLSI Integrated Circuits - A (very) deep dive into computing chips".
- O. Sentieys gave a Keynote at the IEEE International Nanodevices and Computing (INC) - IEEE International Conference on Rebooting Computing (ICRC), Grenoble, France, April 2019 on "Playing with numbers for Energy Efficiency: Introduction to Approximate Computing" [20].
- O. Sentieys gave a tutorial at the 22nd IEEE/ACM Design, Automation and Test in Europe (DATE), March 2019 on "A Comprehensive Analysis of Approximate Computing Techniques: From Component- to Application-Level" [46].
- D.Chillet gave a talk during GDR SoC2 and RO topic day, in november 2019 on "Mathematics model for wavelength allocation in the context of Optical NoC".
- D. Chillet gave an invited talk at FETCH (École d'hiver Francophone sur les Technologies de Conception des Systèmes embarqués Hétérogènes), Louvain-la-Neuve, Belgique, January 2019 on "Power management for communications on Optical NoC".

8.1.5. Leadership within the Scientific Community

- E. Casseau is a member of the French National University Council in Signal Processing and Electronics (CNU - Conseil National des Universites, 61ème section) since 2018.
- D.Chillet is a member of the French National University Council in Signal Processing and Electronics (CNU - Conseil National des Universites, 61ème section) since 2019.
- D. Chillet is member of the Board of Directors of Gretsi Association.
- D. Chillet is co-animator of the "Connected Objects" topic of GDR SoC^2.
- F. Charot and O. Sentieys are members of the steering committee of a CNRS Spring School for graduate students on embedded systems architectures and associated design tools (ARCHI).
- O. Sentieys is a member of the steering committee of GDR SoC^2.
- O. Sentieys is an elected member of the Evaluation Committee (CE) of Inria.

8.1.6. Scientific Expertise

- O. Sentieys served as an expert for Fund for Scientific Research – FNRS, Belgium.
- D.Chillet served as an expert for the call CAPES-COFECUB 2020.
- D.Chillet served as an expert for German Research Foundation.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching Responsibilities

- E. Casseau is in charge of the Department of “Digital Systems” at ENSSAT Engineering Graduate School.
- D. Chillet was the responsible of the ICT Master of University of Science and Technology of Hanoi.
- C. Killian was the responsible of the second year of the ”Instrumentation” DUT at IUT, Lannion until July 2019.
- O. Sentieys is responsible of the ”Embedded Systems” major of the SISEA Master by Research.
- C. Wolinski was the Director of ESIR until May 2019.

ENSSAT stands for *”École Nationale Supérieure des Sciences Appliquées et de Technologie”* and is an *”École d’Ingénieurs”* of the University of Rennes 1, located in Lannion. ISTIC is the Electrical Engineering and Computer Science Department of the University of Rennes 1. ESIR stands for *”École supérieure d’ingénieur de Rennes”* and is an *”École d’Ingénieurs”* of the University of Rennes 1, located in Rennes.

8.2.2. Teaching

E. Casseau: signal processing, 21h, ENSSAT (L3)

E. Casseau: low power design, 6h, ENSSAT (M1)

E. Casseau: real time design methodology, 57h, ENSSAT (M1)

E. Casseau: computer architecture, 24h, ENSSAT (M1)

E. Casseau: VHDL design, 42h, ENSSAT (M1)

E. Casseau: SoC and high-level synthesis, 33h, Master by Research (SISEA) and ENSSAT (M2)

S. Derrien, optimizing and parallelising compilers, 14h, Master of Computer Science, ISTIC(M2)

S. Derrien, advanced processor architectures, 8h, Master of Computer Science, ISTIC(M2)

S. Derrien, high level synthesis, 20h, Master of Computer Science, ISTIC(M2)

S. Derrien, computer science research projects, 10h, Master of Computer Science, ISTIC(M1)

S. Derrien: introduction to operating systems, 8h, ISTIC (M1)

S. Derrien, principles of digital design, 20h, Bachelor of EE/CS, ISTIC(L2)

S. Derrien, computer architecture, 48h, Bachelor of Computer Science, ISTIC(L3)

S.I. Filip, Operating Systems, 24h, Master of Mechatronics, ENS RENNES (M2)

F. Charot: computer architecture, 16h, ESIR (L3)

F. Charot: Computer architecture, 58h, ISTIC (L3)

D. Chillet: embedded processor architecture, 20h, ENSSAT (M1)

D. Chillet: multimedia processor architectures, 24h, ENSSAT (M2)

D. Chillet: advanced processor architectures, 20h, ENSSAT (M2)

D. Chillet: micro-controller, 64h, ENSSAT (L3)

D. Chillet: low-power digital CMOS circuits, 6h, Telecom Bretagne (M2)

D. Chillet: low-power digital CMOS circuits, 4h, UBO (M2)

C. Killian: digital electronics, 52h, IUT Lannion (L1)

C. Killian: automated measurements, 44h, IUT Lannion (L2)

A. Kritikakou: computer architecture 1, 32h, ISTIC (L3)

A. Kritikakou: computer architecture 2, 44h, ISTIC (L3)

A. Kritikakou: C and unix programming languages, 102h, ISTIC (L3)

A. Kritikakou: operating systems, 60h, ISTIC (L3)

O. Sentieys: VLSI integrated circuit design, 24h, ENSSAT (M1)

O. Sentieys: VHDL and logic synthesis, 18h, ENSSAT (M1)

C. Wolinski: computer architectures, 92h, ESIR (L3)

C. Wolinski: design of embedded systems, 48h, ESIR (M1)

C. Wolinski: signal, image, architecture, 26h, ESIR (M1)

C. Wolinski: programmable architectures, 10h, ESIR (M1)

C. Wolinski: component and system synthesis, 10h, Master by Research (ISTIC) (M2)

8.2.3. Supervision

PhD: Audrey Lucas, Software support resistant to passive and active attacks for asymmetric cryptography on (very) small computation cores, Dec. 2019, A. Tisserand.

PhD: Genevieve Ndour, Approximate computing for high energy-efficiency in internet-of-things applications, Jul. 2019, A. Tisserand, A. Molnos (CEA LETI).

PhD in progress: Thibault Allenet, Low-Cost Neural Network Algorithms and Implementations for Temporal Sequence Processing, March 2019, O. Sentieys, O. Bichler (CEA LIST).

PhD in progress: Minh Thanh Cong, Hardware Accelerated Simulation of Heterogeneous Multicore Platforms, May 2017, F. Charot, S. Derrien.

PhD in progress: Minyu Cui, Energy-Quality-Time Fault Tolerant Task Mapping on Multicore Architectures, Oct. 2018, E. Casseau, A. Kritikakou.

PhD in progress: Petr Dobias, Energy-Quality-Time Fault Tolerant Task Mapping on Multicore Architectures, Oct. 2017, E. Casseau.

PhD in progress: Corentin Ferry, Compiler support for Runtime data compression for FPGA accelerators, Sep. 2019, S. Derrien, T. Yuki and S. Rajopadhye (co-tutelle between Université de Rennes 1 and Colorado State University).

PhD in progress: Adrien Gaonac'h, Test de robustesse des systèmes embarqués par perturbation contrôlée en simulation à partir de plateformes virtuelles, Oct. 2019, D. Chillet, Yves Lhuillier (CEA LIST), Youri Helen (DGA).

PhD in progress: Mael Gueguen, Improving the performance and energy efficiency of complex heterogeneous manycore architectures with on-chip data mining, Nov. 2016, O. Sentieys, A. Termier.

PhD in progress: Van-Phu Ha, Application-Level Tuning of Accuracy, Nov. 2017, T. Yuki, O. Sentieys.

PhD in progress: Jaechul Lee, Energy-Performance Trade-Off in Optical Network-on-Chip, Dec. 2018, D. Chillet, C. Killian.

PhD in progress: Thibaut Marty, Compiler support for speculative custom hardware accelerators, Sep. 2017, T. Yuki, S. Derrien.

PhD in progress: Romain Mercier, Fault Tolerant Network on Chip for Deep Learning Algorithms, Oct. 2018, D. Chillet, C. Killian, A. Kritikakou.

PhD in progress: Louis Narmour, Revisiting memory allocation in tyeh polyhedral model, Sep. 2019, S. Derrien, T. Yuki and S. Rajopadhye (co-tutelle between Université de Rennes 1 and Colorado State University).

PhD in progress: Joel Ortiz Sosa, Study and design of a digital baseband transceiver for wireless network-on-chip architectures, Nov. 2016, O. Sentieys, C. Roland (Lab-STICC).

PhD in progress: Davide Pala, Non-Volatile Processors for Intermittently-Powered Computing Systems, Jan. 2018, O. Sentieys, I. Miro-Panades (CEA LETI).

PhD in progress: Joseph Paturel, Design-space exploration of fault-tolerant multicores, Sep. 2018, O. Sentieys, A. Kritikakou.

PhD in progress: Nicolas Roux, Sensor-aided Non-Intrusive Appliance Load Monitoring: Detecting Activity of Devices through Low-Cost Wireless Sensors, Oct. 2016, O. Sentieys, B. Vrigneau (IRISA/GRANIT).

8.3. Popularization

8.3.1. Articles and contents

O. Sentieys contributed to the Inria white book on scientific reviews and challenges in cybersecurity https://files.inria.fr/dircom/extranet/LB_cybersecurity_WEB.pdf.

8.3.2. Interventions

Members of the team participated in the national science festival *(Fête de la Science)* in Lannion, October, with demonstrations on wireless sensor networks, body sensor network, cryptology and digital circuit design.

9. Bibliography

Major publications by the team in recent years

[1] R. David, S. Pillement, O. Sentieys. *Energy-Efficient Reconfigurable Processsors*, in "Low Power Electronics Design", C. Piguet (editor), Computer Engineering, Vol 1, CRC Press, August 2004, chap. 20

[2] S. Derrien, S. Rajopadhye, P. Quinton, T. Risset. *12*, in "High-Level Synthesis From Algorithm to Digital Circuit", P. Coussy, A. Morawiec (editors), Springer Netherlands, 2008, pp. 215-230, http://dx.doi.org/10.1007/978-1-4020-8588-8

[3] C. Huriaux, A. Courtay, O. Sentieys. *Design Flow and Run-Time Management for Compressed FPGA Configurations*, in "IEEE/ACM Design, Automation and Test in Europe (DATE)", March 2015, https://hal.inria.fr/hal-01089319

[4] J.-M. Jézéquel, B. Combemale, S. Derrien, C. Guy, S. Rajopadhye. *Bridging the Chasm Between MDE and the World of Compilation*, in "Journal of Software and Systems Modeling (SoSyM)", October 2012, vol. 11, n^{o} 4, pp. 581-597 [*DOI :* 10.1007/s10270-012-0266-8], https://hal.inria.fr/hal-00717219

[5] B. Le Gal, E. Casseau, S. Huet. *Dynamic Memory Access Management for High-Performance DSP Applications Using High-Level Synthesis*, in "IEEE Transactions on VLSI Systems", 2008, vol. 16, n^{o} 11, pp. 1454-1464

[6] K. Martin, C. Wolinski, K. Kuchcinski, A. Floch, F. Charot. *Constraint Programming Approach to Reconfigurable Processor Extension Generation and Application Compilation*, in "ACM transactions on Reconfigurable Technology and Systems (TRETS)", June 2012, vol. 5, n^{o} 2, pp. 1-38, http://doi.acm.org/10.1145/2209285.2209289

[7] D. Menard, D. Chillet, F. Charot, O. Sentieys. *Automatic Floating-point to Fixed-point Conversion for DSP Code Generation*, in "Proc. ACM/IEEE CASES", October 2002

[8] D. Menard, O. Sentieys. *Automatic Evaluation of the Accuracy of Fixed-point Algorithms*, in "IEEE/ACM Design, Automation and Test in Europe (DATE-02)", Paris, March 2002

[9] S. Pillement, O. Sentieys, R. David. *DART: A Functional-Level Reconfigurable Architecture for High Energy Efficiency*, in "EURASIP Journal on Embedded Systems (JES)", 2008, pp. 1-13

[10] R. Rocher, D. Menard, O. Sentieys, P. Scalart. *Analytical Approach for Numerical Accuracy Estimation of Fixed-Point Systems Based on Smooth Operations*, in "IEEE Transactions on Circuits and Systems. Part I, Regular Papers", October 2012, vol. 59, n^{o} 10, pp. 2326 - 2339 [*DOI :* 10.1109/TCSI.2012.2188938], http://hal.inria.fr/hal-00741741

[11] C. Wolinski, M. Gokhale, K. McCabe. *A polymorphous computing fabric*, in "IEEE Micro", 2002, vol. 22, n^{o} 5, pp. 56–68

[12] C. Wolinski, K. Kuchcinski, E. Raffin. *Automatic Design of Application-Specific Reconfigurable Processor Extensions with UPaK Synthesis Kernel*, in "ACM Trans. on Design Automation of Elect. Syst.", 2009, vol. 15, n^{o} 1, pp. 1–36, http://doi.acm.org/10.1145/1640457.1640458

Publications of the year

Doctoral Dissertations and Habilitation Theses

[13] G. Ndour. *Approximate computing for high energy-efficiency in internet-of-things applications*, Université Rennes 1, July 2019, https://hal.archives-ouvertes.fr/tel-02292988

Articles in International Peer-Reviewed Journals

[14] S.-I. Filip, A. Javeed, L. N. Trefethen. *Smooth random functions, random ODEs, and Gaussian processes*, in "SIAM Review", February 2019, vol. 61, n^{o} 1, pp. 185-205 [*DOI :* 10.1137/17M1161853], https://hal.inria.fr/hal-01944992

[15] G. Gallin, A. Tisserand. *Generation of Finely-Pipelined GF(P) Multipliers for Flexible Curve based Cryptography on FPGAs*, in "IEEE Transactions on Computers", November 2019, vol. 68, n^{o} 11, pp. 1612-1622 [*DOI :* 10.1109/TC.2019.2920352], https://hal.archives-ouvertes.fr/hal-02141260

[16] L. MO, A. KRITIKAKOU, S. HE. *Energy-Aware Multiple Mobile Chargers Coordination for Wireless Rechargeable Sensor Networks*, in "IEEE internet of things journal", May 2019, pp. 1-13 [*DOI :* 10.1109/JIOT.2019.2918837], https://hal.inria.fr/hal-02140283

[17] L. MO, A. KRITIKAKOU. *Mapping Imprecise Computation Tasks on Cyber-Physical Systems*, in "Peer-to-Peer Networking and Applications", 2019, pp. 1726–1740 [*DOI :* 10.1007/s12083-019-00749-9], https://hal.archives-ouvertes.fr/hal-02397099

[18] L. MO, P. YOU, X. CAO, Y.-Q. SONG, A. KRITIKAKOU. *Event-Driven Joint Mobile Actuators Scheduling and Control in Cyber-Physical Systems*, in "IEEE Transactions on Industrial Informatics", March 2019, pp. 1-13 [*DOI :* 10.1109/TII.2019.2906061], https://hal.inria.fr/hal-02080647

[19] S. REDER, F. KEMPF, H. BUCHER, J. BECKER, P. ALEFRAGIS, N. S. VOROS, S. SKALISTIS, S. DERRIEN, I. PUAUT, O. OEY, T. STRIPF, C. FERDINAND, C. DAVID, P. ULBIG, D. MUELLER, U. DURAK. *Worst-Case Execution-Time-Aware Parallelization of Model-Based Avionics Applications*, in "Journal of Aerospace Information Systems", November 2019, vol. 16, n^{o} 11, pp. 521-533 [*DOI :* 10.2514/1.I010749], https://hal.archives-ouvertes.fr/hal-02383381

Invited Conferences

[20] O. SENTIEYS. *Playing with number for Energy Efficiency, Introduction to Approximate Computing*, in "INC 2019 - IEEE International Nanodevices and Computing", Grenoble, France, IEEE, April 2019, https://hal.inria.fr/hal-02183527

International Conferences with Proceedings

[21] T. CONG, F. CHAROT. *Designing Application-Specific Heterogeneous Architectures from Performance Models*, in "MCSoC 2019 - IEEE 13th International Symposium on Embedded Multicore/Many-core Systems-on-Chip", Singapore, Singapore, October 2019, pp. 1-8, https://hal.inria.fr/hal-02289868

[22] M. DARDAILLON, S. SKALISTIS, I. PUAUT, S. DERRIEN. *Reconciling Compiler Optimizations and WCET Estimation Using Iterative Compilation*, in "RTSS 2019 - 40th IEEE Real-Time Systems Symposium", Hong Kong, China, IEEE, December 2019, pp. 1-13, https://hal.archives-ouvertes.fr/hal-02286164

[23] M. GUEGUEN, O. SENTIEYS, A. TERMIER. *Accelerating Itemset Sampling using Satisfiability Constraints on FPGA*, in "DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, IEEE, March 2019, pp. 1046-1051 [*DOI :* 10.23919/DATE.2019.8714932], https://hal.inria.fr/hal-01941862

[24] V.-P. HA, T. YUKI, O. SENTIEYS. *Towards Generic and Scalable Word-Length Optimization*, in "IEEE/ACM Design Automation and Test in Europe (DATE)", Grenoble, France, March 2020, https://hal.inria.fr/hal-02387232

[25] J. LEE, C. KILLIAN, S. L. BEUX, D. CHILLET. *Approximate nanophotonic interconnects*, in "NOCS 2019 - 13th IEEE/ACM International Symposium on Networks-on-Chip", New York, United States, ACM, October 2019, pp. 1-7 [*DOI :* 10.1145/3313231.3352365], https://hal.archives-ouvertes.fr/hal-02341667

[26] O. MATOUSSI, Y. DURAND, O. SENTIEYS, A. MOLNOS. *Error Analysis of the Square Root Operation for the Purpose of Precision Tuning: a Case Study on K-means*, in "ASAP 2019 - 30th IEEE International Conference on Application-specific Systems, Architectures and Processors", New York, United States, IEEE, July 2019, pp. 1-8, https://hal.inria.fr/hal-02183945

[27] L. MO, A. KRITIKAKOU, O. SENTIEYS. *Approximation-aware Task Deployment on Asymmetric Multicore Processors*, in "DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, IEEE, March 2019, pp. 1513-1518 [*DOI :* 10.23919/DATE.2019.8715077], https://hal.inria.fr/hal-01940358

[28] M. S. MOHAMMADI, M. M. STROUT, T. YUKI, K. CHESHMI, E. DAVIS, M. HALL, M. M. DEHNAVI, P. NANDY, C. OLSCHANOWSKY, A. VENKAT. *Sparse computation data dependence simplification for efficient compiler-generated inspectors*, in "PLDI 2019 - 40th ACM SIGPLAN Conference on Programming Language Design and Implementation", Phoenix, United States, ACM Press, November 2019, pp. 594-609 [*DOI :* 10.1145/3314221.3314646], https://hal.inria.fr/hal-02396761

[29] R. PSIAKIS, A. KRITIKAKOU, O. SENTIEYS, E. CASSEAU. *Run-time Coarse-Grained Hardware Mitigation for Multiple Faults on VLIW Processors*, in "DASIP 2019 - Conference on Design and Architectures for Signal and Image Processing", Montréal, Canada, October 2019, pp. 1-6, https://hal.inria.fr/hal-02344282

[30] R. PSIAKIS, A. KRITIKAKOU, O. SENTIEYS. *Fine-Grained Hardware Mitigation for Multiple Long-Duration Transients on VLIW Function Units*, in "DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, IEEE, March 2019, pp. 976-979 [*DOI :* 10.23919/DATE.2019.8714899], https://hal.inria.fr/hal-01941860

[31] S. ROKICKI, D. PALA, J. PATUREL, O. SENTIEYS. *What You Simulate Is What You Synthesize: Designing a Processor Core from C++ Specifications*, in "ICCAD 2019 - 38th IEEE/ACM International Conference on Computer-Aided Design", Westminster, CO, United States, IEEE, November 2019, pp. 1-8, https://hal.archives-ouvertes.fr/hal-02303453

[32] S. ROKICKI, E. ROHOU, S. DERRIEN. *Aggressive Memory Speculation in HW/SW Co-Designed Machines*, in "DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe", Florence, Italy, IEEE, March 2019, pp. 332-335 [*DOI :* 10.23919/DATE.2019.8715010], https://hal.archives-ouvertes.fr/hal-01941876

[33] S. ROKICKI. *GhostBusters: Mitigating Spectre Attacks on a DBT-Based Processor*, in "DATE 2020 - 23rd IEEE/ACM Design, Automation and Test in Europe", Grenoble, France, March 2020, https://hal.archives-ouvertes.fr/hal-02396631

[34] N. ROUX, B. VRIGNEAU, O. SENTIEYS. *Improving NILM by Combining Sensor Data and Linear Programming*, in "SAS 2019 - IEEE Sensors Applications Symposium", Sophia Antipolis, France, IEEE, March 2019, pp. 1-6 [*DOI :* 10.1109/SAS.2019.8706021], https://hal.inria.fr/hal-02394920

[35] B. ROUXEL, S. SKALISTIS, S. DERRIEN, I. PUAUT. *Hiding Communication Delays in Contention-Free Execution for SPM-Based Multi-Core Architectures*, in "ECRTS 2019 - 31st Euromicro Conference on Real-Time Systems", Stuttgart, Germany, July 2019, pp. 1-24 [*DOI :* 10.4230/LIPIcs.ECRTS.2019.25], https://hal.archives-ouvertes.fr/hal-02190271

[36] S. SKALISTIS, A. KRITIKAKOU. *Timely Fine-grained Interference-sensitive Run-time Adaptation of Time-triggered Schedules*, in "RTSS 2019 - 40th IEEE Real-Time Systems Symposium", Hong Kong, China, IEEE, December 2019, pp. 1-13, https://hal.archives-ouvertes.fr/hal-02316392

[37] J. O. SOSA, O. SENTIEYS, C. ROLAND, C. KILLIAN. *Multi-Carrier Spread-Spectrum Transceiver for WiNoC*, in "NOCS 2019 - 13th IEEE/ACM International Symposium on Networks-on-Chip", New York, United States, ACM, October 2019, pp. 1-2 [*DOI :* 10.1145/3313231.3352373], https://hal.inria.fr/hal-02394890

[38] J. O. SOSA, O. SENTIEYS, C. ROLAND. *Adaptive Transceiver for Wireless NoC to Enhance Multicast/Unicast Communication Scenarios*, in "ISVLSI 2019 - IEEE Computer Society Annual Symposium on VLSI", Miami, United States, IEEE, July 2019, pp. 1-6 [*DOI :* 10.1109/ISVLSI.2019.00111], https://hal.inria.fr/hal-02394902

[39] F. DE DINECHIN, S.-I. FILIP, L. FORGET, M. KUMM. *Table-Based versus Shift-And-Add constant multipliers for FPGAs*, in "ARITH 2019 - 26th IEEE Symposium on Computer Arithmetic", Kyoto, Japan, IEEE, June 2019, pp. 1-8, https://hal.inria.fr/hal-02147078

Conferences without Proceedings

[40] V.-P. HA, T. YUKI, O. SENTIEYS. *Noise Budgeting in Multiple-Kernel Word-Length Optimization*, in "AxC 2019 - 4th Workshop on Approximate Computing", Florence, Italy, March 2019, pp. 1-3, https://hal.inria.fr/hal-02183936

[41] S. ROKICKI, D. PALA, J. PATUREL, O. SENTIEYS. *What You Simulate Is What You Synthesize: Design of a RISC-V Core from C++ Specifications*, in "RISC-V Workshop 2019", Zurich, Switzerland, June 2019, pp. 1-2, https://hal.inria.fr/hal-02394911

[42] B. ROUX, M. GAUTIER, O. SENTIEYS. *Exploration architecturale d'accélérateur pour des architectures multi-coeurs hétérogènes*, in "27ème colloque du Groupement de Recherche en Traitement du Signal et des Images", Lille, France, August 2019, https://hal.archives-ouvertes.fr/hal-02406976

[43] T. YUKI. *The Limit of Polynomials: Implications of Handelman's Theorem for Exploring Schedules*, in "IMPACT 2019 - 9th International Workshop on Polyhedral Compilation Techniques", Valencia, Spain, January 2019, pp. 1-8, https://hal.inria.fr/hal-02397043

Scientific Books (or Scientific Book chapters)

[44] D. MENARD, G. CAFFARENA, J. A. LOPEZ, D. NOVO, O. SENTIEYS. *Fixed-point refinement of digital signal processing systems*, in "Digitally Enhanced Mixed Signal Systems", The Institution of Engineering and Technology, May 2019, n^{o} Chapter 1, pp. 1-37 [*DOI :* 10.1049/PBCS040E_CH], https://hal.inria.fr/hal-01941898

[45] D. MÉNARD, G. CAFFARENA, J. A. LOPEZ, D. NOVO, O. SENTIEYS. *Analysis of Finite Word-Length Effects in Fixed-Point Systems*, in "Handbook of Signal Processing Systems", S. S. BHATTACHARYYA (editor), 2019, pp. 1063-1101 [*DOI :* 10.1007/978-3-319-91734-4_29], https://hal.inria.fr/hal-01941888

Other Publications

[46] A. BOSIO, D. MENARD, O. SENTIEYS. *A Comprehensive Analysis of Approximate Computing Techniques: From Component- to Application-Level*, March 2019, pp. 1-5, DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe, https://hal.inria.fr/hal-01941757

[47] M. KUMM, A. VOLKOVA, S.-I. FILIP. *Design of Optimal Multiplierless FIR Filters*, December 2019, https://arxiv.org/abs/1912.04210 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02392522

[48] S. ROKICKI, E. ROHOU, S. DERRIEN. *Hybrid-DBT: Hardware Accelerated Dynamic Binary Translation*, June 2019, 1 p. , RISC-V 2019 - Workshop Zurich, Poster, https://hal.archives-ouvertes.fr/hal-02155019

[49] Y. Uguen, F. De Dinechin, V. Lezaud, S. Derrien. *Application-specific arithmetic in high-level synthesis tools*, October 2019, working paper or preprint [*DOI :* 10.1145/nnnnnnn.nnnnnnn], https://hal.archives-ouvertes.fr/hal-02178465

References in notes

[50] S. Hauck, A. DeHon (editors). *Reconfigurable Computing: The Theory and Practice of FPGA-Based Computation*, Morgan Kaufmann, 2008

[51] V. Baumgarte, G. Ehlers, F. May, A. Nückel, M. Vorbach, M. Weinhardt. *PACT XPP — A Self-Reconfigurable Data Processing Architecture*, in "The Journal of Supercomputing", 2003, vol. 26, n^{o} 2, pp. 167–184

[52] C. Beckhoff, D. Koch, J. Torresen. *Portable module relocation and bitstream compression for Xilinx FPGAs*, in "24th Int. Conf. on Field Programmable Logic and Applications (FPL)", 2014, pp. 1–8

[53] C. Bobda. *Introduction to Reconfigurable Comp.: Architectures Algorithms and Applications*, Springer, 2007

[54] S. Borkar, A. A. Chien. *The Future of Microprocessors*, in "Commun. ACM", May 2011, vol. 54, n^{o} 5, pp. 67–77, http://doi.acm.org/10.1145/1941487.1941507

[55] J. M. P. Cardoso, P. C. Diniz, M. Weinhardt. *Compiling for reconfigurable computing: A survey*, in "ACM Comput. Surv.", June 2010, vol. 42, 13:1 p. , http://doi.acm.org/10.1145/1749603.1749604

[56] K. Compton, S. Hauck. *Reconfigurable computing: a survey of systems and software*, in "ACM Comput. Surv.", 2002, vol. 34, n^{o} 2, pp. 171–210, http://doi.acm.org/10.1145/508352.508353

[57] J. Cong, H. Huang, C. Ma, B. Xiao, P. Zhou. *A Fully Pipelined and Dynamically Composable Architecture of CGRA*, in "IEEE Int. Symp. on Field-Program. Custom Comput. Machines (FCCM)", 2014, pp. 9–16, http://dx.doi.org/10.1109/FCCM.2014.12

[58] G. Constantinides, P. Cheung, W. Luk. *Wordlength optimization for linear digital signal processing*, in "IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems", October 2003, vol. 22, n^{o} 10, pp. 1432- 1442

[59] M. Coors, H. Keding, O. Luthje, H. Meyr. *Fast Bit-True Simulation*, in "Proc. ACM/IEEE Design Automation Conference (DAC)", Las Vegas, june 2001, pp. 708-713

[60] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, A. R. LeBlanc. *Design of ion-implanted MOSFET's with very small physical dimensions*, in "IEEE Journal of Solid-State Circuits", 1974, vol. 9, n^{o} 5, pp. 256–268

[61] A. Hormati, M. Kudlur, S. Mahlke, D. Bacon, R. Rabbah. *Optimus: efficient realization of streaming applications on FPGAs*, in "Proc. ACM/IEEE CASES", 2008, pp. 41–50

[62] H. Kalte, M. Porrmann. *REPLICA2Pro: Task Relocation by Bitstream Manipulation in Virtex-II/Pro FPGAs*, in "3rd Conference on Computing Frontiers (CF)", 2006, pp. 403–412

[63] J.-E. Lee, K. Choi, N. D. Dutt. *Compilation Approach for Coarse-Grained Reconfigurable Architectures*, in "IEEE Design and Test of Computers", 2003, vol. 20, n^{o} 1, pp. 26-33, http://doi.ieeecomputersociety.org/10.1109/MDT.2003.1173050

[64] H. Lee, D. Nguyen, J.-E. Lee. *Optimizing Stream Program Performance on CGRA-based Systems*, in "52nd IEEE/ACM Design Automation Conference", 2015, pp. 110:1–110:6, http://doi.acm.org/10.1145/2744769.2744884

[65] B. Mei, S. Vernalde, D. Verkest, H. De Man, R. Lauwereins. *ADRES: An architecture with tightly coupled VLIW processor and coarse-grained reconfigurable matrix*, in "Proc. FPL", Springer, 2003, pp. 61–70

[66] N. R. Miniskar, S. Kohli, H. Park, D. Yoo. *Retargetable Automatic Generation of Compound Instructions for CGRA Based Reconfigurable Processor Applications*, in "Proc. ACM/IEEE CASES", 2014, pp. 4:1–4:9, http://doi.acm.org/10.1145/2656106.2656125

[67] Y. Park, H. Park, S. Mahlke. *CGRA express: accelerating execution using dynamic operation fusion*, in "Proc. Int. Conf. on Compilers, Architecture, and Synthesis for Embedded Systems", New York, NY, USA, CASES'09, ACM, 2009, pp. 271–280, http://doi.acm.org/10.1145/1629395.1629433

[68] A. Putnam, A. Caulfield, E. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. Gopal, J. Gray, M. Haselman, S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P. Xiao, D. Burger. *A reconfigurable fabric for accelerating large-scale datacenter services*, in "ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)", June 2014, pp. 13-24, http://dx.doi.org/10.1109/ISCA.2014.6853195

[69] G. Theodoridis, D. Soudris, S. Vassiliadis. *2*, in "A survey of coarse-grain reconfigurable architectures and CAD tools", Springer Verlag, 2007

[70] G. Venkataramani, W. Najjar, F. Kurdahi, N. Bagherzadeh, W. Bohm, J. Hammes. *Automatic compilation to a coarse-grained reconfigurable system-on-chip*, in "ACM Trans. on Emb. Comp. Syst.", 2003, vol. 2, n^{o} 4, pp. 560–589, http://doi.acm.org/10.1145/950162.950167