RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

CNRS, Université Rennes 1, École normale supérieure de Rennes

2020

ACTIVITY REPORT

Project-Team

CAIRN

Energy Efficient Computing ArchItectures with Embedded Reconfigurable Resources

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

DOMAIN

Algorithmics, Programming, Software and Architecture

THEME

Architecture, Languages and Compilation

Contents

Project-Team CAIRN

Creation of the Project-Team: 2009 January 01

Keywords

Computer sciences and digital sciences

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

A1.6. – Green Computing

A2.2. – Compilation

A2.2.1. – Static analysis

A2.2.4. – Parallel architectures

A2.2.6. – GPGPU, FPGA...

A2.2.7. – Adaptive compilation

A2.3.1. – Embedded systems

A2.3.3. – Real-time systems

A4.4. – Security of equipment and software

A8.10. – Computer arithmetic

A9. – Artificial intelligence

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 members, visitors, external collaborators

Research Scientists

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

Faculty Members

- Olivier Sentieys [Team leader, Univ de Rennes I, Professor, Inria Chair, HDR]
- Emmanuel Casseau [Univ de Rennes I, Professor, HDR]
- Daniel Chillet [Univ de Rennes I, Professor, HDR]
- Steven Derrien [Univ de Rennes I, Professor, HDR]
- Cédric Killian [Univ de Rennes I, Associate Professor]
- Angeliki Kritikakou [Univ de Rennes I, Associate Professor]
- Patrice Quinton [École normale supérieure de Rennes, Emeritus]
- Christophe Wolinski [Univ de Rennes I, Professor, HDR]

Post-Doctoral Fellows

- Joel Ortiz Sosa [Inria]
- Simon Rokicki [École normale supérieure de Rennes]

PhD Students

- Thibault Allenet [CEA]
- Minh Thanh Cong [Univ de Rennes I, until Sep 2020]
- Minyu Cui [China Scholarship Council]
- Petr Dobias [Univ de Rennes I, until Aug 2020]
- Corentin Ferry [Univ de Rennes I]
- Adrien Gaonac H [CEA]
- Cedric Gernigon [Inria, from Oct 2020]
- Van Phu Ha [Inria]
- Ibrahim Krayem [Univ de Rennes I, from Oct 2020]
- Jaechul Lee [Univ de Rennes I]
- Thibaut Marty [Univ de Rennes I, until Nov 2020]
- Romain Mercier [Inria]
- Davide Pala [Inria]
- Joseph Paturel [Inria, until Sep 2020]
- Yuxiang Xie [Inria, from Oct 2020]

Technical Staff

- Justine Bonnot [Inria, Engineer, until Oct 2020]
- Pierre Halle [Univ de Rennes I, Engineer]
- Mickaël Le Gentil [Univ de Rennes I, Engineer]

Interns and Apprentices

- Chloe Briand [Inria, from May 2020 until Aug 2020]
- Leo Cosseron [Univ de Rennes I, from May 2020 until Jul 2020]
- Lauric Desauw [Inria, from May 2020 until Aug 2020]
- Cedric Gernigon [Univ de Rennes I, from Apr 2020 until Sep 2020]
- Timothee Kocev [Univ de Rennes I, from May 2020 until Aug 2020]
- Tom Malherbe [Inria, from Jul 2020 until Aug 2020]
- Louis Savary [Univ de Rennes I, from May 2020 until Jul 2020]

Administrative Assistants

- Emilie Carquin [Univ de Rennes I]
- Nadia Derouault [Inria]

Visiting Scientist

- Jinyi Xu [China Scholarship Council, from Nov 2020]

2 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) [49] and then by Coarse-Grain Reconfigurable Arrays (CGRA) [41, 54] [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 [44], 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 [48], 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 [42]. 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. 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).

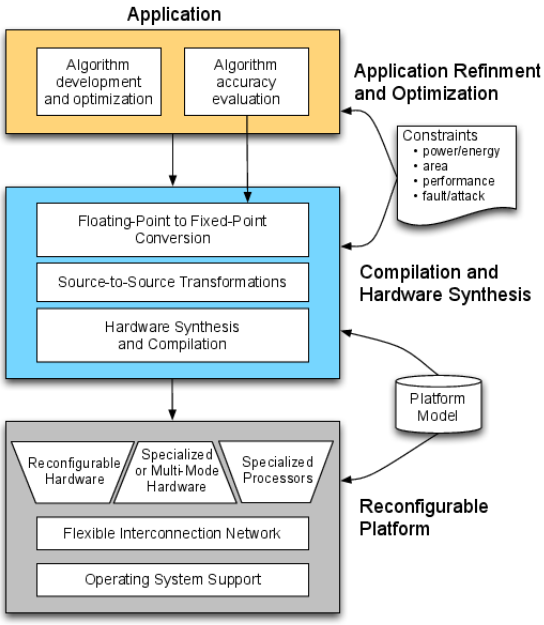

Figure 1: CAIRN's general design flow and related research themes

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 [57]. They can also be reconfigured dynamically. At run-time, 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 [40] or implementing an online filter to relocate the tasks [51]. 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 [58, 45, 56]. 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 [54] or polymorphous computing fabrics[11]. These works have led to

commercial products such as the PACT/XPP [39] 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 [43, 59, 55, 52, 53]. 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 [50]. 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) [3], 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 [5].

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 [46]. 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 [47]. 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

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 New software

5.1.1 Gecos

Name: 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.

URL: http://gecos.gforge.inria.fr

Author: Steven Derrien

Contact: Steven Derrien

Participants: Tomofumi Yuki, Thomas Lefeuvre, Imèn Fassi, Mickael Dardaillon, Ali Hassan El Moussawi, Steven Derrien

Partner: Université de Rennes 1

5.1.2 ID-Fix

Name: 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 in 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 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 optimized 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 constructs an analytic accuracy model of the program, which means it can explore more solutions and thereby produce a much more efficient code.

URL: http://idfix.gforge.inria.fr

Authors: Daniel Menard, Olivier Sentieys, Loïc Cloatre, Nicolas Simon, Quentin Meunier, Jean-Charles Naud, Romuald Rocher

Contact: Olivier Sentieys

Participant: Olivier Sentieys

Partner: Université de Rennes 1

5.1.3 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

5.2 New platforms

5.2.1 Zyggie: a Wireless Body Sensor Network Platform

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: Mickael Le Gentil and Olivier Sentieys
- Partners: Lab-STICC, Université de Rennes 1
- Contact: Olivier Sentieys
- URL: https://project.inria.fr/bowi/zyggie-wbsn-platform

Figure 2: CAIRN's Zyggie platform for WBSN

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

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

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

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

5.2.4 Hybrid-DBT

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 enables an overall performance increase of 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

5.2.5 Comet

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

5.2.6 TypEx

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

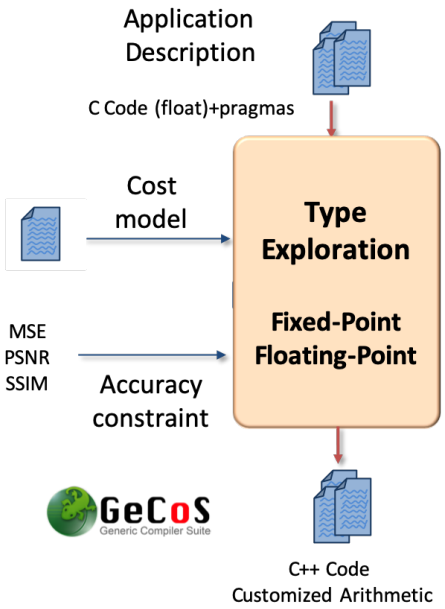

Figure 3: TypEx: a tool for type exploration and automatic floating-point to fixed-point conversion

6 New results

6.1 Freezer: A Specialized NVM Backup Controller for Intermittently-Powered Systems

Participants Davide Pala, Olivier Sentieys.

The explosion of IoT and wearable devices generated a rising attention towards energy harvesting as source for powering these systems. In this context, many applications cannot afford the presence of a battery because of size, weight and cost issues. Therefore, due to the intermittent nature of ambient energy sources, these systems must be able to save and restore their state, in order to guarantee progress across power interruptions. In [17], we propose a specialized backup/restore controller that dynamically tracks the memory accesses during the execution of the program. The controller then commits the changes to a snapshot in a Non-Volatile Memory (NVM) when a power failure is detected. Our approach does not require complex hybrid memories and can be implemented with standard components. Results

on a set of benchmarks show an average 8× reduction in backup size. Thanks to our dedicated controller, the backup time is further reduced by more than 100×, with an area and power overhead of only 0.4% and 0.8%, respectively, w.r.t. a low-end IoT node.

6.2 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.

In [15], we proposed a technique for improving the efficiency of convolutional neural network hardware accelerators based on timing speculation (overclocking) and algorithmic fault tolerance (ABFT). We augment the accelerator with a lightweight error detection mechanism to protect against timing errors in convolution layers, enabling aggressive timing speculation. The error detection mechanism we have developed works at the algorithm level, utilizing algebraic properties of the computation, allowing the full implementation to be realized using high-level synthesis tools. Our prototype on ZC706 demonstrated up to 60% higher throughput with negligible area overhead for various wordlength implementations.

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 extended precision is not much more costly than implementing the computation in standard floating point. Our current work revolves around evaluating the approach when used for ABFT.

6.3 Speculative pipelining for High Level Synthesis

Participants Steven Derrien, Simon Rokicki, Thibaut Marty, Tomofumi Yuki.

High-Level Synthesis (HLS) aims at making hardware design more accessible by enabling the automatic derivation of highly optimized hardware designs, directly from high-level specifications (C, C++, OpenCL). Although HLS is now a mature technology, existing tools still suffer from many limitations, and are usually less efficient compared to manual designs by experts. HLS tools take algorithmic specifications in the form of C/C++ as inputs, just like standard compilers. They hence benefit from decades of research in optimizing program transformations, such as loop pipelining (also known as modulo sheduling). Loop pipelining enables the synthesis of complex yet area-efficient hardware datapaths. Current HLS tools are very effective at applying loop pipelining to loops with regular control and simple memory access patterns, but struggle with data-dependent control-flow and/or memory accesses. The main reason is that existing loop pipelining techniques rely on static schedules, which cannot precisely capture data-dependent behaviors.

We have started addressing this limitation by studying how Loop Pipelining can be extended to support dynamic and speculative behavior [13]. Unlike prior work on the topic, we express speculation directly at the program level. This allows for a seamless integration into HLS design flows, and provides two key benefits: (i) the pipelined datapath is synthesized by the HLS tools that are capable of deriving

efficient designs, and (ii) we do not compromise on the ease-of-use aspects: programmers keep all the productivity benefits (e.g., easier/faster validation) of having high-level specifications. The technique is being implemented within a source to source compiler [13], and our first results show significant performance improvements over standard loop pipelining for many kernels.

6.4 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 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.

6.5 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. Reasoning 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 [23], we describe a methodology allowing to explore the design space of power-performance heterogeneous SoCs by combining an architecture simulator (gem5-Aladdin) and a hyperparameter optimization method (Hyperopt). This methodology allows different types of parallelism with loop unrolling strategies and memory coherency interfaces to be swept. It has been applied to a convolutional neural network algorithm. We show that the most energy efficient architecture achieves a 2× to 4× improvement in energy-delay-product compared to an architecture without parallelism. Furthermore, the obtained solution is more efficient than commonly implemented architectures (Systolic, 2D-mapping, and Tiling). We also applied the methodology to find the optimal architecture including its coherency interface for a complex SoC made up of six accelerated-workloads. We show that a hybrid interface appears to be the most efficient; it reaches 22% and 12% improvement in energy-delay-product compared to using only non-coherent and only LLC-coherent models, respectively.

6.6 Fault-Tolerant Scheduling onto Multicore embedded Systems

Participants Emmanuel Casseau, Minyu Cui, Petr Dobias, Angeliki Kritikakou.

We have still considered how to map and schedule tasks onto homogeneous faulty processors. Two approaches have been investigated. The first approach deals with task mapping onto processors before runtime. We studied the problem of task mapping by jointly solving task allocation, task frequency assignment, task duplication decision on multicore platforms enhanced with DVFS capability [24]. The goal is to minimize energy consumption without violating real-time constraints, taking also reliability into account. The problem is initially formulated as Integer Non-Linear Programming and equivalently transformed to a Mixed Integer Linear Programming problem to be optimally solved. This work is done in collaboration with Lei Mo, School of Automation, Southeast University (China). The second approach deals with mapping and scheduling tasks at runtime. The application context is CubeSat nanosatellites. Cubestats have to respect time, spatial and energy constraints in harsh space environment. We proposed two fault tolerant online scheduling algorithms [25, 26]. The first algorithm considers all tasks as aperiodic tasks and the second one treats them as aperiodic or periodic tasks. The choice is subject to a trade-off between number of processors, rejection rate and energy consumption. This work is done in collaboration with Oliver Sinnen, PARC Lab., the University of Auckland (New-Zealand).

6.7 Run-time management of Real-Time safety-critical systems

Participants Angeliki Kritikakou.

Worst-Case Execution Time (WCET) estimations are required, during deployment of real-time safety-critical systems, in order to provide guarantees for real-time execution. However, static estimations of WCET are safe, but over-estimated, leading to over-provisioned systems and underutilized cores. To solve this problem, we propose run-time adaptation approaches to exploit the information that becomes available as the execution progresses. Our first dynamic approach safely adapts time-triggered schedules, obtained offline with interference-sensitive WCET [33]. This is achieved by relaxing or completely removing schedule dependencies, exploiting the progress of each core. In this way, an earlier task execution is enabled, creating time slack that can be used by safety-critical and mixed-criticality systems to provide higher Quality-of-Services or execute other best-effort applications. We also presented the response time analysis of the proposed approach , showing that although the approach is dynamic, it is fully predictable with bounded WCET. The approach has been evaluated for different application types and execution configurations on the 8-core Texas Instruments TMS320C6678 platform, obtaining significant performance improvements compared to static approaches. Our second approach focuses on mixed-critical systems, consisting of high criticality and low criticality applications. Typically, when a high criticality task exceeds its less pessimistic WCET, the system switches mode and low criticality tasks are usually dropped. Our approach exposes the slack created due to the progress of tasks, during execution, and safely uses it in order to postpone mode switch.

6.8 Energy Constrained and Real-Time Imprecise Computation Tasks Mapping on Networked Systems

Participants Olivier Sentieys, Angeliki Kritikakou.

Networked systems are useful for a wide range of applications, many of which require distributed and collaborative data processing to satisfy real-time requirements. On the one hand, networked systems are usually resource constrained, mainly regarding the energy supply of the nodes and their computation and communication abilities. On the other hand, many real-time applications can be executed in an imprecise way, where an approximate result is acceptable as long as the baseline Quality-of-Service (QoS) is satisfied. Such applications can be modeled through Imprecise Computation (IC) tasks. To achieve a better trade-off between QoS and limited system resources, while meeting application requirements, the IC-tasks must be efficiently mapped to the system nodes. To tackle this problem, in [16], we construct an IC-tasks mapping problem that aims to maximize system QoS subject to real-time and energy constraints.

Dynamic Voltage and Frequency Scaling (DVFS) and multi-path routing are explored to further enhance real-time performance and reduce energy consumption. Secondly, based on the problem structure, we propose an optimal approach to perform IC-tasks mapping and prove its optimality. Furthermore, to enhance the scalability of the proposed approach, we present a heuristic IC-tasks mapping method with low computation time. Finally, the simulation results demonstrate the effectiveness of the proposed methods in terms of the solution quality and the computation time.

6.9 Fault-Tolerant Microarchitectures

Participants Joseph Paturel, Angeliki Kritikakou, Olivier Sentieys.

Due to technology scaling and harsh environments, a wide range of fault-tolerant techniques exists to deal with the error occurrences. Selecting a fault-tolerant technique is not trivial, whereas more than the necessary overhead is usually inserted during the system design. To avoid over-designing, it is necessary to have an in-depth understanding of the available design options. However, an exhaustive listing is neither possible to create nor efficient to use due to its prohibitive size. In [14], we present a top-down binary tree classification for error detection and correction techniques. At each split, the design space is clearly divided into two complementary parts using one single attribute, compared with existing classifications that use splits with multiple attributes. A leaf inherits all the attributes of its ancestors from the root to the leaf. A technique is decomposed into primitive components, each one belonging to a different leaf. The single attribute splits can be used to efficiently compare the techniques and to prune the incompatible parts of the design space during the design of a technique. This essential single attribute division of the design space is required for the improvement of the techniques and for novel contributions to the fault-tolerance domain.

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. In [31], we propose 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.

6.10 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), or at a high hardware cost. For that, 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. In [29], we evaluated the proposed method for data mining benchmark and we show that our proposal can lead to a reduction from 10^{-2} to 10^{-8} for the centroid mean position Mean Square Error (MSE) in the K-means clustering algorithm with a limited area cost and power consumption.

6.11 Optical Network-on-Chip

Participants Jaechul Lee, Cédric Killian, Joel Ortiz Sosa, Daniel Chillet.

The energy consumption of manycore is dominated by data movements, which calls for energy-efficient and high-bandwidth interconnects. Classical solutions, based on electrical NoC, suffer from low scalability which lead to performance reduction. 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. To exploit the main characteristics of optical NoC, we develop an approximate communication model for data exchanges based on laser power management. The data to transfer are classified into sensitive data and data which can be approximated without to much Quality of Service (QoS) degradation. From this classification, we are able to reduce the energy of communication by reducing the laser power of LSB bits (Least Significant Bits) and/or by truncating them, while the MSB bits are sent at nominal power level. The SNR of LSB is then reduced or truncated impacting the communication QoS. We have developed a simulation platform, based on Sniper, and we demonstrate 53% of power saving and 8% of output errors for the StreamCluster application from Approxbench suite.

Furthermore, we also explored the Crosstalk Power Penalty in the context of Wavelength Division Multiplexing (WDM) communications on ONoC. This technique allows several wavelengths to be allocated for a communication to speed up data transfers at the cost of power penalty, due to crosstalk involved by the MicroRing Resonators at each receiver. We have shown that, for 16 wavelengths allocated for a communication, the power penalty can vary from 1.2 dB for the highest wavelength, to 5 dB for the wavelengths located in the middle of the bandwidth. Nonetheless, we also explored the power penalty with respect to the number of wavelengths used among the maximum available for a communication. As an example, we have shown that for 16 available wavelengths, the power penalty varies from 0 dB for only one wavelength allocated, to a range from 1.5dB to 4.5dB for 8 allocated wavelengths on the 16 available. Furthermore, we proposed a mathematical formalization (based on OPL language and Cplex solver) of design space to evaluate the two extreme bounds of design spaces (execution time vs power consumption) considering application and architecture parameters. We show that, for generic applications, the execution time can vary from 43% and energy from 30% between the two extreme bounds [34].

6.12 Word-Length Optimization

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

Using just the right amount of numerical precision is an important aspect for guaranteeing performance and energy efficiency requirements. Word-Length Optimization (WLO) is the automatic process for tuning the precision, i.e., bit-width, of variables and operations represented using fixed-point arithmetic. However, state-of-the-art precision tuning approaches do not scale well in large applications where many variables are involved.

In [28], 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 [27], we propose a hybrid algorithm combining Bayesian optimization (BO) and a fast local search to speed up the WLO procedure. Through experiments, we first show some evidence on how this combination can improve exploration time. Then, we propose an algorithm to automatically determine a

reasonable transition point between the two algorithms. By statistically analyzing the convergence of the probabilistic models constructed during BO, we derive a stopping condition that determines when to switch to the local search phase. Experimental results indicate that our algorithm can reduce exploration time by up to 50%-80% for large benchmarks.

6.13 Application-specific arithmetic in high-level synthesis tools

Participants Steven Derrien.

In [19], 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.

7 Partnerships and cooperations

7.1 International initiatives

7.1.1 Inria Associate Team

IntelliVIS

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

Duration: 2019 - 2022

Coordinator: Olivier Sentieys

Partners: IIT Goa (India)

Inria contact: Olivier Sentieys

Summary: 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 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 would be an interesting and challenging research direction.

7.1.2 Inria international partners

DARE

Title: Design space exploration Approaches for Reliable Embedded systems

Partners: IMEC (Belgium) - Francky Catthoor, IMEC fellow

Inria contact: Angeliki Kritikakou

Summary: 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.

LRS

Title: Loop unRolling Stones: compiling in the polyhedral model

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

Inria contact: Steven Derrien

HARAMCOP

Title: Hardware accelerators modeling using constraint-based programming

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

Inria contact: Christophe Wolinski

DeLeES

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

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

Inria contact: Olivier Sentieys

Summary: 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.

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
- School of Informatics, Aristotle University of Thessaloniki (Greece), Memory management, fault tolerance
- Raytheon Technologies, Ireland, run-time management for time-critical systems
- Karlsruhe Institute of Technology - KIT (Germany), Loop parallelization and compilation techniques for embedded multicores.
- PARC Lab., Department of Electrical, Computer, and Software Engineering, the University of Auckland (New-Zealand), Fault-tolerant task scheduling onto multicore.
- Ruhr - University of Bochum - RUB (Germany), Reconfigurable architectures.
- School of Automation, Southeast University (China), Fault-tolerant task scheduling onto multi-core.
- Shantou University (China), Runtime efficient algorithms for subgraph enumeration.
- University of Science and Technology of Hanoi (Vietnam), Participation of several CAIRN's members in the Master ICT / Embedded Systems.

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. 2023
- 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 synchronizing the accesses to shared data, requires to distribute and share information between some 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 Regional initiatives

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://cominlabs.inria.fr

8 Dissemination

8.1 Promoting scientific activities

8.1.1 Scientific events: organisation

General chair, scientific chair

- D. Chillet was the General Co-Chair of HiPEAC RAPIDO'20 Workshop.
- D. Chillet will be the General Chair of ARC'21.

Member of the organizing committees

- O. Sentieys will serve as Local Chair of the Organizing Committee of ARC'21.
- A. Kritikakou will serve as Publicity and Web Chair of ARC'21.
- A. Kritikakou served as Publication Co-Chair of ISVLSI'20 conference.
- A. Kritikakou will serve as Artifact evaluation Co-Chair of ECRTS'21 and ECRTS'22 conference.
- C. Killian will serve in Organizing Committee of ARC'21.

8.1.2 Scientific events: selection

Chair of conference program committees

- O. Sentieys was Co-Chair of the D8 Track on Architectural and Microarchitectural Design at IEEE/ACM DATE from 2018 to 2020.
- O. Sentieys is Chair of the D9 Track on Architectural and Microarchitectural Design at IEEE/ACM DATE 2021.
- O. Sentieys was 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).
- S. Derrien will serve as Program Chair of ARC'21.

Member of the conference program committees

- E. Casseau was a member of the technical program committee of FPT.
- 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, IEEE ASAP, ICPP and ARC.
- A. Kritikakou was a member of technical program committee of IEEE RTAS, ECRTS, SAMOS, HPPC, IISWC and DATE.
- O. Sentieys was a member of technical program committee of IEEE/ACM DATE, IEEE FPL, ACM ENSSys, ACM SBCCI, IEEE ReConFig, and FDL.
- T. Yuki was a member of technical program committee of CGO, SC, TAPAS, CC, and IMPACT.

8.1.3 Journal

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.
- A. Kritikakou was a member of the editorial board of the Special Issue on "Holistic Technologies for Sustainable Cyber-Physical Systems", Microprocessors and Microsystems, Elsevier
- A. Kritikakou was a member of the editorial board of the Special Issue on the Special Issue on "Software and Hardware Co-Design for Sustainable Cyber-Physical Systems", Journal of Software : Practice and Experience, Wiley Press
- A. Kritikakou was a member of the editorial board of the Special Issue on "Cyber-Physical System Technologies for Sustainable Power and Energy Systems", CSEE Journal of Power and Energy Systems (CSEE JPES)

8.1.4 Leadership within the scientific community

- 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.

8.1.5 Scientific expertise

- O. Sentieys served as an expert for: Fund for Scientific Research (FNRS) of Belgium and Natural Sciences and Engineering Research Council of Canada (NSERC).
- O. Sentieys is a member of the ANR Scientific Evaluation Committee CE25 "Software science and engineering - Multi-purpose communication networks, high-performance infrastructure".
- D.Chillet served as an expert for IRP (International Research Project) of INS2I dept of CNRS.
- D.Chillet served as an expert for HCERES evaluation.

8.1.6 Research administration

- S. Derrien is the head of the D3 "Architecture" Department of IRISA.

8.2 Teaching - Supervision

8.2.1 Teaching Responsibilities

- E. Casseau is in charge of the Department of “Digital Systems” at ENSSAT Engineering Graduate School.
- D. Chillet is the new responsible of the ”Embedded Systems” major of the SISEA Master by Research since september 2020.
- C. Killian is the responsible of the second year of the ”Instrumentation” DUT at IUT, Lannion.
- O. Sentieys was responsible of the ”Embedded Systems” major of the SISEA Master by Research until august 2020.
- 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, 42h, ESIR (L3)
- F. Charot: Computer architecture, 44h, ISTIC (L3)
- F. Charot: Compilation and code optimization architecture, 18h, ENSSAT (M2)
- 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, 4h, UBO (M2)
- C. Killian: digital electronics, 52h, IUT Lannion (L1)
- C. Killian: automated measurements, 44h, IUT Lannion (L2)
- C. Killian: computer architecture, 4h, IUT Lannion (L3)
- 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 PhD Supervision

- PhD: Petr Dobias, Energy-Quality-Time Fault Tolerant Task Mapping on Multicore Architectures, Oct. 2020, E. Casseau.
- PhD: Mael Gueguen, Frequent Itemset Sampling of High Throughput Streams on FPGA Accelerators, Oct. 2020, O. Sentieys, A. Termier (Lacodam).
- PhD: Joel Ortiz Sosa, Design of a Digital Baseband Transceiver for Wireless Network-on-Chip Architectures, Nov. 2020, O. Sentieys, C. Roland (Lab-STICC).
- 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: 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: Cédric Gernigon, Highly compressed/quantized neural networks for FPGA onboard processing in Earth observation by satellite, Oct. 2020, O. Sentieys, S. Filip.
- PhD in progress: Van-Phu Ha, Application-Level Tuning of Accuracy, Nov. 2017, T. Yuki, O. Sentieys.
- PhD in progress: Ibrahim Krayem, Fault tolerant emerging on-chip interconnects for manycore architectures, Oct. 2020, C. Killian, D. Chillet.
- 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: Davide Pala, Non-Volatile Processors for Intermittently-Powered Computing Systems, Jan. 2018, O. Sentieys, I. Miro-Panades (CEA LETI).
- PhD in progress: Leo Pradels, Constrained optimization of FPGA accelerators for embedded deep convolutional neural networks, Dec. 2020, D. Chillet, O. Sentieys, S. Filip.
- PhD in progress: Yuxiang Xie, Efficient Low-Precision Training for Deep Learning Accelerators, Oct. 2020, O. Sentieys.

8.3 Popularization

8.3.1 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, and digital circuit design.

9 Scientific production

9.1 Major publications

[1] R. David, S. Pillement and O. Sentieys. 'Energy-Efficient Reconfigurable Processsors'. In: *Low Power Electronics Design*. Ed. by C. Piguet. Computer Engineering, Vol 1. CRC Press, Aug. 2004. Chap. 20.

[2] S. Derrien, S. Rajopadhye, P. Quinton and T. Risset. 'High-Level Synthesis From Algorithm to Digital Circuit'. In: ed. by P. Coussy and A. Morawiec. Springer Netherlands, 2008. Chap. 12, pp. 215–230. DOI: 10.1007/978-1-4020-8588-8. URL: http://dx.doi.org/10.1007/978-1-4020-8588-8.

[3] B. L. Gal, E. Casseau and S. Huet. 'Dynamic Memory Access Management for High-Performance DSP Applications Using High-Level Synthesis'. In: *IEEE Transactions on VLSI Systems* 16.11 (2008), pp. 1454–1464.

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

[5] J.-M. Jézéquel, B. Combemale, S. Derrien, C. Guy and S. Rajopadhye. 'Bridging the Chasm Between MDE and the World of Compilation'. In: *Journal of Software and Systems Modeling (SoSyM)* 11.4 (Oct. 2012), pp. 581–597. DOI: 10.1007/s10270-012-0266-8. URL: https://hal.inria.fr/hal-00717219.

[6] K. Martin, C. Wolinski, K. Kuchcinski, A. Floch and F. Charot. 'Constraint Programming Approach to Reconfigurable Processor Extension Generation and Application Compilation'. In: *ACM transactions on Reconfigurable Technology and Systems (TRETS)* 5.2 (June 2012), pp. 1–38. DOI: 10.1145/2209285.2209289. URL: http://doi.acm.org/10.1145/2209285.2209289.

[7] D. Menard, D. Chillet, F. Charot and O. Sentieys. 'Automatic Floating-point to Fixed-point Conversion for DSP Code Generation'. In: *Proc. ACM/IEEE CASES*. Oct. 2002.

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

[9] S. Pillement, O. Sentieys and 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 and 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* 59.10 (Oct. 2012), pp. 2326–2339. DOI: 10.1109/TCSI.2012.2188938. URL: http://hal.inria.fr/hal-00741741.

[11] C. Wolinski, M. Gokhale and K. McCabe. 'A polymorphous computing fabric'. In: *IEEE Micro* 22.5 (2002), pp. 56–68.

[12] C. Wolinski, K. Kuchcinski and E. Raffin. 'Automatic Design of Application-Specific Reconfigurable Processor Extensions with UPaK Synthesis Kernel'. In: *ACM Trans. on Design Automation of Elect. Syst.* 15.1 (2009), pp. 1–36. URL: http://doi.acm.org/10.1145/1640457.1640458.

9.2 Publications of the year

International journals

[13] S. Derrien, T. Marty, S. Rokicki and T. Yuki. 'Toward Speculative Loop Pipelining for High-Level Synthesis'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.11 (2020). URL: https://hal.archives-ouvertes.fr/hal-02949516.

[14] A. Kritikakou, R. Psiakis, F. Catthoor and O. Sentieys. 'Binary Tree Classification of Rigid Error Detection and Correction Techniques'. In: *ACM Computing Surveys* 53.4 (25th Aug. 2020), pp. 1–38. DOI: 10.1145/3397268. URL: https://hal.archives-ouvertes.fr/hal-02927439.

[15] T. Marty, T. Yuki and S. Derrien. 'Safe Overclocking for CNN Accelerators through Algorithm-Level Error Detection'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (Mar. 2020), pp. 4777–4790. DOI: 10.1109/TCAD.2020.2981056. URL: https://hal.inria.fr/hal-03094811.

[16] L. Mo, A. Kritikakou, O. Sentieys and X. Cao. 'Real-time Imprecise Computation Tasks Mapping for DVFS-Enabled Networked Systems'. In: *IEEE internet of things journal* (17th Dec. 2020), p. 13. DOI: 10.1109/JIOT.2020.3044910. URL: https://hal.archives-ouvertes.fr/hal-03103821.

[17] D. Pala, I. Miro-Panades and O. Sentieys. 'Freezer: A Specialized NVM Backup Controller for Intermittently-Powered Systems'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2020), pp. 1–15. DOI: 10.1109/TCAD.2020.3025063. URL: https://hal.inria.fr/hal-03119369.

[18] B. Roux, M. Gautier, O. Sentieys and J.-P. Delahaye. 'Energy-driven design space exploration of tiling-based accelerators for heterogeneous multiprocessor architectures'. In: *Microprocessors and Microsystems: Embedded Hardware Design (MICPRO)* 77 (2020), pp. 1–12. DOI: 10.1016/j.micpro.2020.103138. URL: https://hal.inria.fr/hal-02747622.

[19] Y. Uguen, F. De Dinechin, V. Lezaud and S. Derrien. 'Application-specific arithmetic in high-level synthesis tools'. In: *ACM Transactions on Architecture and Code Optimization* (1st Mar. 2020). DOI: 10.1145/3377403. URL: https://hal.archives-ouvertes.fr/hal-02423363.

[20] C. Xiao, S. Wang, W. Liu, X. Wang and E. Casseau. 'An Optimal Algorithm for Enumerating Connected Convex Subgraphs in Acyclic Digraphs'. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 68.1 (5th Jan. 2021), pp. 261–265. DOI: 10.1109/TCSII.2020.3000297. URL: https://hal.inria.fr/hal-02884025.

International peer-reviewed conferences

[21] N. Bellec, S. Rokicki and I. Puaut. 'Attack detection through monitoring of timing deviations in embedded real-time systems'. In: ECRTS 2020 - 32nd Euromicro Conference on Real-Time Systems. Modena, Italy, 7th July 2020, pp. 1–22. DOI: 10.4230/LIPIcs.ECRTS.2020.8. URL: https://hal.inria.fr/hal-02559549.

[22] J. Bonnot, D. Menard and K. Desnos. 'Fast Kriging-based Error Evaluation for Approximate Computing Systems'. In: Design, Automation & Test in Europe Conference & Exhibition (DATE). Grenoble, France, 9th Mar. 2020. URL: https://hal.archives-ouvertes.fr/hal-02468086.

[23] T. Cong and F. Charot. 'Design Space Exploration of Heterogeneous-Accelerator SoCs with Hyperparameter Optimization'. In: 26th Asia and South Pacific Design Automation Conference (ASP-DAC 2021). Virtual Conference, Japan, 18th Jan. 2021. URL: https://hal.inria.fr/hal-03119732.

[24] M. Cui, L. Mo, A. Kritikakou and E. Casseau. 'Energy-aware Partial-Duplication Task Mapping under Real-Time and Reliability Constraints'. In: SAMOS 2020 - International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation. Samos / Virtual, Greece, 5th July 2020. URL: https://hal.inria.fr/hal-02927474.

[25] P. Dobiáš, E. Casseau and O. Sinnen. 'Evaluation of the Fault-Tolerant Online Scheduling Algorithms for CubeSats'. In: DSD 2020 - 23rd EUROMICRO Conference on Digital System Design. Portoroz, Slovenia, 26th Aug. 2020, pp. 1–11. URL: https://hal.inria.fr/hal-02927553.

[26] P. Dobiáš, E. Casseau and O. Sinnen. 'Fault-Tolerant Online Scheduling Algorithms for CubeSats'. In: PARMA-DITAM'20 - 11th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architecture, 9th Workshop on Design Tools and Architectures for Multicore Embedded Computing Platforms. Bologna, Italy, 21st Jan. 2020. URL: https://hal.inria.fr/hal-02461164.

[27] V.-P. Ha and O. Sentieys. 'Leveraging Bayesian Optimization to Speed Up Automatic Precision Tuning'. In: 24th IEEE/ACM Design, Automation and Test in Europe (DATE). Virtual Event, France, 2021. URL: https://hal.inria.fr/hal-03119548.

[28] V.-P. Ha, T. Yuki and O. Sentieys. 'Towards Generic and Scalable Word-Length Optimization'. In: DATE 2020 - 23rd IEEE/ACM Design, Automation and Test in Europe. Grenoble, France, 9th Mar. 2020, pp. 1–6. URL: https://hal.inria.fr/hal-02387232.

[29] R. Mercier, C. Killian, A. Kritikakou, Y. Helen and D. Chillet. 'Multiple Permanent Faults Mitigation Through Bit-Shuffling for Network-on-Chip Architecture'. In: ICCD 2020 - IEEE International Conference on Computer Design. Hartford / Virtual, United States, 18th Oct. 2020. URL: https://hal.inria.fr/hal-03039545.

[30] V. L. Nguyen Huu, J. Lallet, E. Casseau and L. D'orazio. 'MASCARA (ModulAr Semantic CAching fRAmework) towards FPGA acceleration for IoT Security monitoring'. In: VLIoT 2020 - International Workshop on Very Large Internet of Things. Vol. 6. Open Journal of Internet Of Things (OJIOT) 1. Tokyo, Japan: http://nbn-resolving.de/urn:nbn:de:101:1-202008021933291279842 6, 4th Sept. 2020, pp. 14–23. URL: https://hal.inria.fr/hal-03017402.

[31] J. Paturel, A. Kritikakou and O. Sentieys. 'Fast Cross-Layer Vulnerability Analysis of Complex Hardware Designs'. In: ISVLSI 2020 - IEEE Computer Society Annual Symposium on VLSI. Limassol, Cyprus: https://www.computer.org/conferences/cfp/ISVLSI2020, 6th July 2020, pp. 328–333. DOI: 10.1109/ISVLSI49217.2020.00067. URL: https://hal.archives-ouvertes.fr/hal-02927455.

[32] S. Rokicki. 'GhostBusters: Mitigating Spectre Attacks on a DBT-Based Processor'. In: *DATE 2020 - 23rd IEEE/ACM Design, Automation and Test in Europe*. DATE 2020 - 23rd IEEE/ACM Design, Automation and Test in Europe. Grenoble, France, 9th Mar. 2020, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-02396631.

[33] S. Skalistis and A. Kritikakou. 'Dynamic Interference-Sensitive Run-time Adaptation of Time-Triggered Schedules'. In: ECRTS 2020 - 32nd Euromicro Conference on Real-Time Systems. ECRTS 2020 - 32nd Euromicro Conference on Real-Time Systems. Virtual, France, 7th July 2020, pp. 1–22. DOI: 10.4230/LIPIcs.ECRTS.2020.4. URL: https://hal.archives-ouvertes.fr/hal-02927451.

[34] J. O. Sosa, C. Killian, H. B. Ammar and D. Chillet. 'Min/max time limits and energy penalty of communication scheduling in ring-based ONoC'. In: NoCArc 2020 - 13th International Workshop on Network on Chip Architectures. On-line, France, 18th Oct. 2020. URL: https://hal.inria.fr/hal-03032687.

Conferences without proceedings

[35] A. Kritikakou and S. Skalistis. 'Progress-aware Dynamic Slack Exploitation in Mixed-critical Systems: Work-in-Progress'. In: EMSOFT 2020 - International Conference on Embedded Software. Hamburg / Virtual, Germany: https://sigbed.org/emsoft-2020/, 20th Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-03125812.

Doctoral dissertations and habilitation theses

[36] P. Dobiáš. 'Online fault tolerant task scheduling for real-time multiprocessor embedded systems'. Université Rennes 1, 2nd Oct. 2020. URL: https://hal.archives-ouvertes.fr/tel-03016351.

[37] M. Gueguen. 'Frequent Itemset Sampling of High Throughput Streams on FPGA Accelerators'. Université de Rennes 1 (UR1), 23rd Oct. 2020. URL: https://tel.archives-ouvertes.fr/tel-03120148.

[38] J. O. Sosa. 'Design of a Digital Baseband Transceiver for Wireless Network-on- Chip Architectures'. Université de Rennes 1 (UR1), 11th Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03120129.

9.3 Cited publications

[39] V. Baumgarte, G. Ehlers, F. May, A. Nückel, M. Vorbach and M. Weinhardt. 'PACT XPP — A Self-Reconfigurable Data Processing Architecture'. In: *The Journal of Supercomputing* 26.2 (2003), pp. 167–184.

[40] C. Beckhoff, D. Koch and 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.

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

[42] S. Borkar and A. A. Chien. 'The Future of Microprocessors'. In: *Commun. ACM* 54.5 (May 2011), pp. 67–77. DOI: 10.1145/1941487.1941507. URL: http://doi.acm.org/10.1145/1941487.1941507.

[43] J. M. P. Cardoso, P. C. Diniz and M. Weinhardt. 'Compiling for reconfigurable computing: A survey'. In: *ACM Comput. Surv.* 42 (4 June 2010), 13:1. DOI: http://doi.acm.org/10.1145/1749603.1749604. URL: http://doi.acm.org/10.1145/1749603.1749604.

[44] K. Compton and S. Hauck. 'Reconfigurable computing: a survey of systems and software'. In: *ACM Comput. Surv.* 34.2 (2002), pp. 171–210. DOI: http://doi.acm.org/10.1145/508352.508353. URL: http://doi.acm.org/10.1145/508352.508353.

[45] J. Cong, H. Huang, C. Ma, B. Xiao and 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. DOI: 10.1109/FCCM.2014.12. URL: http://dx.doi.org/10.1109/FCCM.2014.12.

[46] G. Constantinides, P. Cheung and W. Luk. 'Wordlength optimization for linear digital signal processing'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 22.10 (Oct. 2003), pp. 1432–1442.

[47] M. Coors, H. Keding, O. Luthje and H. Meyr. 'Fast Bit-True Simulation'. In: *Proc. ACM/IEEE Design Automation Conference (DAC)*. Las Vegas, June 2001, pp. 708–713.

[48] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous and A. R. LeBlanc. 'Design of ion-implanted MOSFET's with very small physical dimensions'. In: *IEEE Journal of Solid-State Circuits* 9.5 (1974), pp. 256–268.

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

[50] A. Hormati, M. Kudlur, S. Mahlke, D. Bacon and R. Rabbah. 'Optimus: efficient realization of streaming applications on FPGAs'. In: *Proc. ACM/IEEE CASES*. Atlanta, GA, USA, 2008, pp. 41–50.

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

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

[53] J.-E. Lee, K. Choi and N. D. Dutt. 'Compilation Approach for Coarse-Grained Reconfigurable Architectures'. In: *IEEE Design and Test of Computers* 20.1 (2003), pp. 26–33. URL: http://doi.ieeecomputersociety.org/10.1109/MDT.2003.1173050.

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

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

[56] Y. Park, H. Park and S. Mahlke. 'CGRA express: accelerating execution using dynamic operation fusion'. In: *Proc. Int. Conf. on Compilers, Architecture, and Synthesis for Embedded Systems*. CASES'09. Grenoble, France: ACM, 2009, pp. 271–280. DOI: http://doi.acm.org/10.1145/1629395.1629433. URL: http://doi.acm.org/10.1145/1629395.1629433.

[57] 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 and 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. DOI: 10.1109/ISCA.2014.6853195. URL: http://dx.doi.org/10.1109/ISCA.2014.6853195.

[58] G. Theodoridis, D. Soudris and S. Vassiliadis. 'A survey of coarse-grain reconfigurable architectures and CAD tools'. In: *Fine- and coarse-grain reconfigurable computing*. Springer Verlag, 2007. Chap. 2.

[59] G. Venkataramani, W. Najjar, F. Kurdahi, N. Bagherzadeh, W. Bohm and J. Hammes. 'Automatic compilation to a coarse-grained reconfigurable system-on-chip'. In: *ACM Trans. on Emb. Comp. Syst.* 2.4 (2003), pp. 560–589. URL: http://doi.acm.org/10.1145/950162.950167.