UMR IRISA

Activity Report 2021

Team TARAN

Domain-Specific Computers in the Post Moore's Law Era

Joint team with Inria Rennes – Bretagne Atlantique

D3 – Architecture

CentraleSupélec · cnrs · rennes · IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · Inria · INSA RENNES · Université Bretagne Sud ubs · UNIVERSITÉ DE RENNES 1

Contents

Project-Team TARAN 1

1 Team members, visitors, external collaborators 2

2 Overall objectives 4
- 2.1 Context: End of CMOS . . . 4
- 2.2 Design Stack for Custom Hardware . . . 4
- 2.3 Objectives of TARAN: Facilitating Cross-Layer Optimization . . . 5

3 Research program 5
- 3.1 Accelerators . . . 6
- 3.2 Accurate Computing . . . 6
- 3.3 Resilient Computing . . . 6
- 3.4 Embracing Emerging Technologies . . . 7

4 Application domains 7

5 New software and platforms 8
- 5.1 New software . . . 8
  - 5.1.1 Gecos . . . 8
  - 5.1.2 SmartSense . . . 8
  - 5.1.3 TypEx . . . 9
- 5.2 New platforms . . . 9
  - 5.2.1 E-methodHW: an automatic tool for the evaluation of polynomial and rational function approximations . . . 9
  - 5.2.2 Firopt: a tool for the simultaneous design of digital FIR filters along with the dedicated hardware model . . . 9
  - 5.2.3 Hybrid-DBT . . . 10
  - 5.2.4 Comet . . . 10

6 New results 10
- 6.1 Improving Memory Throughput of Hardware Accelerators . . . 10
- 6.2 High-Level Synthesis of Speculative Hardware Accelerators . . . 11
- 6.3 Design Space Exploration for IoT Processors Platforms . . . 11
- 6.4 Hardware Accelerated Simulation of Heterogeneous Platforms . . . 11
- 6.5 Recursive Polyhedral Equations . . . 12
- 6.6 Energy Constrained and Real-Time Imprecise Computation Tasks Mapping on Networked Systems . . . 12
- 6.7 Training Deep Neural Networks with Low-Precision Accelerators . . . 13
- 6.8 Word-Length Optimization . . . 13
- 6.9 Towards an Arithmetic-Centered Global Approach for Filter Design . . . 14
- 6.10 Fault-Tolerant Microarchitectures . . . 14
- 6.11 Fault-Tolerant Networks-on-Chip . . . 14
- 6.12 Fault-Tolerant Task Deployment onto Multicore Systems . . . 15
- 6.13 Freezer: A Specialized NVM Backup Controller for Intermittently-Powered Systems . . . 15
- 6.14 Optical Network-on-Chip for error resilient applications . . . 15
- 6.15 Dynamic Optical Network-on-Chip based Phase Change Material . . . 16

7 Bilateral contracts and grants with industry 16
- 7.1 Bilateral contracts with industry . . . 16
- 7.2 Bilateral Grants with Industry . . . 16
- 7.3 Informal Collaborations with Industry . . . 17

8 Partnerships and cooperations **17**
8.1 International initiatives . . . 17
8.1.1 Inria Associate Team . . . 17
8.1.2 Inria International Partners . . . 17
8.2 International research visitors . . . 18
8.2.1 Visits of international scientists . . . 18
8.2.2 Visits to international teams . . . 18
8.3 National initiatives . . . 19
8.3.1 ANR AdequateDL . . . 19
8.3.2 ANR RAKES . . . 19
8.3.3 ANR Opticall2 . . . 20
8.3.4 ANR SHNOC . . . 20
8.3.5 DGA RAPID - FLODAM (2017–2021) . . . 21
8.3.6 ANR FASY . . . 21
8.3.7 ANR Re-Trusting . . . 21
8.3.8 DGA/INRIA Sniffer . . . 22
8.3.9 Labex CominLabs - LeanAI (2021-2024) . . . 22

9 Dissemination **23**
9.1 Promoting scientific activities . . . 23
9.1.1 Scientific events: organisation . . . 23
9.1.2 Scientific events: selection . . . 23
9.1.3 Journal . . . 23
9.1.4 Invited talks . . . 24
9.1.5 Leadership within the scientific community . . . 24
9.1.6 Scientific expertise . . . 24
9.1.7 Research administration . . . 24
9.2 Teaching - Supervision . . . 25
9.2.1 Teaching Responsibilities . . . 25
9.2.2 Teaching . . . 25
9.2.3 PhD Supervision . . . 26

10 Scientific production **27**
10.1 Major publications . . . 27
10.2 Publications of the year . . . 28
10.3 Cited publications . . . 30

Project-Team TARAN

Creation of the Project-Team: 2021 May 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

A2.2. – Compilation

A2.2.4. – Parallel architectures

A2.2.6. – GPGPU, FPGA...

A2.2.7. – Adaptive compilation

A2.2.8. – Code generation

A2.3.1. – Embedded systems

A2.3.3. – Real-time systems

A4.4. – Security of equipment and software

A8.10. – Computer arithmetic

A9.9. – Distributed AI, Multi-agent

Other research topics and application domains

B4.5. – Energy consumption

B4.5.1. – Green computing

B4.5.2. – Embedded sensors consumption

B6.4. – Internet of things

B6.6. – Embedded systems

1 Team members, visitors, external collaborators

Research Scientists

- François Charot [Inria, Researcher, from May 2021]
- Silviu Filip [Inria, Researcher, from May 2021]
- Tomofumi Yuki [Inria, Researcher, from May 2021 until Jul 2021]

Faculty Members

- Olivier Sentieys [Team leader, Univ de Rennes I, Professor, from May 2021, HDR]
- Emmanuel Casseau [Univ de Rennes I, Professor, from May 2021, HDR]
- Daniel Chillet [Univ de Rennes I, Professor, from May 2021, HDR]
- Steven Derrien [Univ de Rennes I, Professor, from May 2021, HDR]
- Cédric Killian [Univ de Rennes I, Associate Professor, from May 2021]
- Angeliki Kritikakou [Univ de Rennes I, Associate Professor, from May 2021]
- Patrice Quinton [École normale supérieure de Rennes, Emeritus, from May 2021]
- Simon Rokicki [École normale supérieure de Rennes, Associate Professor, from Sep 2021]
- Christophe Wolinski [Univ de Rennes I, Professor, from May 2021 until Aug 2021, HDR]

Post-Doctoral Fellows

- Yash Agrawal [Univ de Rennes I, from May 2021]
- Sonia Barrios Pereira [Inria, from May 2021]
- Abhijit Das [Univ de Rennes I, from Nov 2021]
- Fernando Fernandes Dos Santos [Inria, from Nov 2021]
- Marcello Traiola [Inria, from Oct 2021]

PhD Students

- Thibault Allenet [CEA, from May 2021]
- Minyu Cui [China Scholarship Council, from May 2021]
- Léo De La Fuente [CEA, From Dec 2021]
- Corentin Ferry [Univ de Rennes I, from May 2021]
- Adrien Gaonac'h [CEA, from May 2021]
- Cedric Gernigon [Inria, from May 2021]
- Jean Michel Gorius [Univ de Rennes I, from Sep 2021]
- Ibrahim Krayem [Univ de Rennes I, from May 2021]
- Seungah Lee [Univ de Rennes 1, From Nov 2021]
- Jaechul Lee [Univ de Rennes I, from May 2021 until Nov 2021]

- Amelie Marotta [Inria, from Oct 2021]
- Romain Mercier [Inria, from May 2021]
- Leo Pradels [Groupe SAFRAN, CIFRE, from May 2021]
- Yuxiang Xie [Inria, from May 2021 until Sep 2021]

Technical Staff

- Logan Fortune [Inria, Engineer, from May 2021 until Jul 2021]
- Pierre Halle [Inria, Engineer, from May 2021]
- Mickaël Le Gentil [Univ de Rennes I, Engineer, from May 2021 until Aug 2021]
- Arash Nejat [Inria, Engineer, from May 2021]
- Joel Ortiz Sosa [Univ de Rennes I, Engineer, from May 2021]
- Simon Rokicki [École normale supérieure de Rennes, Engineer, from Jan 2021 until Aug 2021]

Interns and Apprentices

- Abdelrahman Ahmed Mahmoud Azab Ali [Inria, from May 2021 until Aug 2021]
- Hind Ait Taleb [Univ de Rennes I, from Jun 2021 until Aug 2021]
- Nitesh Narayana Gondlyala Sathya [Univ de Rennes I, from Jun 2021 until Aug 2021]
- Jean Michel Gorius [École normale supérieure de Rennes, from May 2021 until Jul 2021]
- Mathis Lavigne [Inria, from Jun 2021 until Aug 2021]
- Anthony Le Guyader [Univ de Rennes I, from Jun 2021 until Jul 2021]
- Dylan Leothaud [Univ de Rennes I, from May 2021 until Jul 2021]
- Stefan Locke-Robin [Univ de Rennes I, from May 2021]
- Thomas Mevel [Univ de Rennes I, from Jun 2021 until Aug 2021]
- Mouad Moatassim Billah [Univ de Rennes I, from May 2021 until Sep 2021]
- Remi Robilliard [Univ de Rennes I, from May 2021 until Jul 2021]
- Antoine Solcourt [École normale supérieure de Rennes, from May 2021 until Aug 2021]

Administrative Assistants

- Emilie Carquin [Univ de Rennes I, from May 2021]
- Nadia Derouault [Inria, from May 2021]

Visiting Scientist

- Jinyi Xu [China Scholarship Council, from Jun 2021]

2 Overall objectives

Energy efficiency has now become one of the main requirements for virtually all computing platforms [51]. We now have an opportunity to refine our objectives in order to address the computing challenges of the next couple of decades, with the most prominent one being the end of CMOS scaling. Our belief is that the key to sustaining improvements in performance (both speed and energy) is *domain-specific computing* where all layers of computing, from languages and compilers to runtime and circuit design, must be carefully tailored to specific contexts.

2.1 Context: End of CMOS

Few years ago, the Dennard scaling was starting to breakdown [50, 49], posing new challenges around energy and power consumption. We are now at the end of another important trend in computing, Moore's Law, that brings another set of challenges.

Moore's Law is Running Out of Steam The limits of traditional transistor process technology have been known for a long time. We are now approaching these limits while alternative technologies are still in early stages of development. The economical drive for more performance will persist, and we expect a surge in specialized architectures in the mid-term to squeeze performance out of CMOS technology. Use of Non-Volatile Memory (NVM), Processing-in-Memory (PIM), and various work on approximate computing are all examples of such architectures.

Specialization is the Common Denominator Specialization, which has been a small niche in the past, is now widespread [46]. The main driver today is energy efficiency—small embedded devices need specialized hardware to operate under power/energy constraints. In the next ten years, we expect specializations to become even more common to meet increasing demands for performance. In particular, high-throughput workloads traditionally ran on servers (e.g., computational science and machine learning) will offload (parts of) their computations to accelerators. We are already seeing some instances of such specialization, most notably accelerators for neural networks that use clusters of nodes equipped with FPGAs and/or ASICs.

The Need for Abstractions The main drawback of hardware specialization is that it comes with significant costs in terms of productivity. Although High-Level Synthesis tools have been steadily improving, design and implementation of custom hardware (HW) are still time consuming tasks that require significant expertise. As specializations become inevitable, we need to provide programmers with tools to develop specialized accelerators and explore their large design spaces. Raising the level of abstraction is a promising way to improve productivity, but also introduces additional challenges to maintain the same levels of performance as manually specified counterparts. Taking advantage of domain knowledge to better automate the design flow from higher level specifications to efficient implementations is necessary for making specialized accelerators accessible.

2.2 Design Stack for Custom Hardware

We view the custom hardware design stack as the five layers described below. Our core belief is that next-generation architectures require the expertise in these layers to be efficiently combined.

Language/Programming Model This is the main interface to the programmer that has two (sometimes conflicting) goals. One is that the programmer should be able to concisely specify the computation. The other is that the domain knowledge of the programmer must also be expressed such that the other layers can utilize it.

Compiler The compiler is an important component for both productivity and performance. It improves productivity by allowing the input language to be more concise by recovering necessary information through compiler analysis. It is also where the first set of analyses and transformations are performed to realize efficient custom hardware.

Runtime Runtime complements adjacent layers with its dynamicity. It has access to more concrete information about the input data that static analyses cannot use. It is also responsible for coordinating various processing elements, especially in heterogeneous settings.

Hardware Design There are many design knobs when building an accelerator: the amount/type of parallelism, communication and on-chip storage, number representation and computer arithmetic, and so on. The key challenge is in navigating through this design space with the help of domain knowledge passed through the preceding layers.

Emerging Technology Use of non-conventional hardware components (e.g., NVM or optical interconnects) opens further avenues to explore specialized designs. For a domain where such emerging technologies make sense, this knowledge should also be taken into account when designing the HW.

2.3 Objectives of TARAN: Facilitating Cross-Layer Optimization

Our main objective is to promote Domain-Specific Computing that requires the participation of the algorithm designer, the compiler writer, the microarchitect, and the chip designer. This cannot happen through individually working on the different layers discussed above. The unique composition of TARAN allows us to benefit from our expertise spanning multiple layers in the design stack.

3 Research program

Our research directions may be categorized into the following four contexts:

- **Accelerators**: Hardware accelerators will become more and more common, and we must develop techniques to make accelerator design more accessible. The important challenge is raising the level of abstraction without sacrificing performance. However, higher level of abstraction coupled with domain-specific knowledge is also a great opportunity to widen the scope of accelerators.

- **Accurate Computing**: Most computing today is performed with significant over-provisioning of output quality or precision. Carefully selecting the various parameters, ranging from algorithms to arithmetic, to compute with just the right quality is necessary for further efficiency. Such fine tuning of elements affecting application quality is extremely time consuming and requires domain knowledge to be fully utilized.

- **Resilient Computing**: As we approach the limit of CMOS scaling, it becomes increasingly unlikely for a computing device to be fully functional due to various sources of faults. Thus, techniques to maintain efficiency in the presence of faults will be important. Generally applicable techniques, such as replication, come with significant overheads. Developing techniques tailored to each application will be necessary for computing contexts where reliability is critical.

- **Embracing Emerging Technologies**: Certain computing platforms, such as ultra-low power devices and embedded many-cores, have specific design constraints that make traditional components unfit. However, emerging technologies such as Non-Volatile Memory and Silicon Photonics cannot simply be used as a substitute. Effectively integrating more recent technologies is an important challenge for these specialized computing platforms.

The common keyword across all directions is **domain-specific**. Specialization is necessary for addressing various challenges including productivity, efficiency, reliability, and scalability in the next generation of computing platforms. Our main objective is defined by the need to jointly work on multiple layers of the design stack to be truly domain-specific. Another common challenge for the entire team is **design space exploration**, which has been and will continue to be an essential process for HW design. We can only expect the design space to keep expanding, and we must persist on developing techniques to efficiently navigate through the design space.

3.1 Accelerators

Key Investigators: E. Casseau, F. Charot, D. Chillet, S. Derrien, A. Kritikakou, P. Quinton, O. Sentieys.
Accelerators are custom hardware that primarily aim to provide high-throughput, energy-efficient, computing platforms. Custom hardware can give much better performance compared to more general architectures simply because they are specialized, at the price of being much harder to "program." Accelerator designers need to explore a massive design space, which includes many hardware parameters that a software programmer has no control over, to find a suitable design for the application at hand.

Our first objective in this context is to further enlarge the design space and enhance the performance of accelerators. The second, equally important, objective is to provide the designers with the means to efficiently navigate through the ever-expanding design space. Cross-layer expertise is crucial in achieving these goals—we need to fully utilize available domain knowledge to improve both the productivity and the performance of custom hardware design.

Positioning Hardware acceleration has already proved its efficiency in many datacenter, cloud-computing or embedded high-performance computing (HPC) applications: machine learning, web search, data mining, database access, information security, cryptography, financial, image/signal/video processing, etc. For example, the work at Microsoft in accelerating the Bing web search engine with large-scale reconfigurable fabrics has shown to improve the ranking throughput of each server by 95% [55], and the increasing need for acceleration of deep learning workloads [58].

Hardware accelerators still lack efficient and standardized compilation toolflows, which makes the technology impractical for large-scale use. Generating and optimizing hardware from high-level specifications is a key research area with considerable interest [47, 53]. On this topic, 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.

3.2 Accurate Computing

Key Investigators: S. Filip, S. Derrien, O. Sentieys.
An important design knob in accelerators is the number representation—digital computing is by nature some approximation of real world behavior. Appropriately selecting the number representation that respects a given quality requirement has been a topic of study for many decades in signal/image processing: a process known as Word-Length Optimization (WLO). We are now seeing the scope of number format-centered approximations widen beyond these traditional applications. This gives us many more approximation opportunities to take advantage of, but introduces additional challenges as well.

Earlier work on arithmetic optimizations has primarily focused on low-level representations of the computation (i.e., signal-flow graphs) that do not scale to large applications. Working on higher level abstractions of the computation is a promising approach to improve scalability and to explore high-level transformations that affect accuracy. Moreover, the acceptable degree of approximation is decided by the programmer using domain knowledge, which needs to be efficiently utilized.

Positioning Traditionally, fixed-point (FxP) arithmetic is used to relax accuracy, providing important benefits in terms of delay, power and area [15]. There is also a large body of work on carefully designing efficient arithmetic operators/functions that preserve good numerical properties. Such numerical precision tuning leads to a massive design space, necessitating the development of efficient and automatic exploration methods.

The need for further improvements in energy efficiency has led to renewed interest in approximation techniques in the recent years [54]. This field has emerged in the last years, and is very active recently with deep learning as its main driver. Many applications have modest numerical accuracy requirements, allowing for the introduction of approximations in their computations [48].

3.3 Resilient Computing

Key Investigators: E. Casseau, D. Chillet, C. Killian, A. Kritikakou, O. Sentieys.

With advanced technology nodes and the emergence of new devices pressured by the end of Moore's law, manufacturing problems and process variations strongly influence electrical parameters of circuits and architectures [52], leading to dramatically reduced yield rates [56]. Transient errors caused by particles or radiations will also more and more often occur during execution [59, 57], and process variability will prevent predicting chip performance (e.g., frequency, power, leakage) without a self-characterization at run time. On the other hand, many systems are under constant attacks from intruders and security has become of utmost importance.

In this research direction, we will explore techniques to protect architectures against faults, errors, and attacks, which have not only a low overhead in terms of area, performance, and energy [17, 16, 12], but also a significant impact on improving the resilience of the architecture under consideration. Such protections require to act at most layers of the design stack.

3.4 Embracing Emerging Technologies

Key Investigators: D. Chillet, S. Derrien, C. Killian, O. Sentieys.

Domain specific accelerators have more exploratory freedom to take advantage of non-conventional technologies that are too specialized for general purpose use. Examples of such technologies include optical interconnects for Network-on-Chip (NoC) and Non-Volatile Memory (NVM) for low-power sensor nodes. The objective of this research direction is to explore the use of such technologies, and find appropriate application domains. The primary cross-layer interaction is expected from Hardware Design to accommodate non-conventional Technologies. However, this research direction may also involve Runtime and Compilers.

4 Application domains

Application Domains Spanning from Embedded Systems to Datacenters Computing systems are the invisible key enablers for all Information and Communication Technologies (ICT) innovations. Until recently, computing systems were mainly hidden under a desk or in a machine room. But future efficient computing systems should embrace different application domains, from sensors or smartphones to cloud infrastructures. The next generation of computer systems are facing enormous challenges. The computer industry is in the midst of a major shift in how it delivers performance because silicon technologies are reaching many of their power and performance limits. Contributing to post Moore's law domain-specific computers will have therefore significant societal impact in almost all application domains.

In addition to recent and widespread portable devices, new embedded systems such as those used in medicine, robots, drones, etc., already demand high computing power with stringent constraints on energy consumption, especially when implementing computationally-intensive algorithms, such as the now widespread inference and training of Deep Neural Networks (DNNs). As examples, we will work on defining efficient computing architectures for DNN inference on resource-constrained embedded systems (e.g., on-board satellite, IoT devices), as well as for DNN training on FPGA accelerators or on edge devices.

The class of applications that benefit from hardware accelerations has steadily grown over the past years. Signal processing and image processing are classic examples which are still relevant. Recent surge of interest towards deep learning has led to accelerators for machine learning (e.g., Tensor Processing Units). In fact, it is one of our tasks to expand the domain of applications amenable to acceleration by reducing the burden on the programmers/designers. We have recently explored accelerating Dynamic Binary Translation [19] and we will continue to explore new application domains where HW acceleration is pertinent.

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: https://gitlab.inria.fr/gecos

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 SmartSense

Name: Sensor-Aided Non-Intrusive Load Monitoring

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.

URL: https://smartsense.inria.fr/

Contact: Olivier Sentieys

5.1.3 TypEx

Name: Type Exploration Tool

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

Scientific Description: 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 wordlengths is the one found by the tool that respects the accuracy constraint given and that minimizes a parametrized cost function.

Functional 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. TypEx is available open-source at https://gitlab.inria.fr/gecos/gecos-float2fix. See README.md for detailed instructions on how to install the software.

URL: https://gitlab.inria.fr/gecos/gecos-float2fix

Contact: Olivier Sentieys

5.2 New platforms

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

KEYWORDS: function approximation, FPGA hardware implementation generator

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

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

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

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

KEYWORDS: FIR filter design, multiplierless hardware implementation generator

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

- Participants: Silviu-Ioan Filip, Martin Kumm, Anastasia Volkova
- Partners: Univ Rennes, Université de Nantes, Fulda University of Applied Sciences

- Contact: Silviu-Ioan Filip
- URL: https://gitlab.com/filteropt/firopt

5.2.3 Hybrid-DBT

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

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

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

5.2.4 Comet

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

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

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

6 New results

6.1 Improving Memory Throughput of Hardware Accelerators

Participants Steven Derrien, Corentin Ferry, Tomofumi Yuki.

Offloading compute-intensive kernels to hardware accelerators relies on the large degree of parallelism offered by these platforms. However, the effective bandwidth of the memory interface often causes a bottleneck, hindering the accelerator's effective performance. Techniques enabling data reuse, such as tiling, lower the pressure on memory traffic but still often leave the accelerators I/O-bound. A further increase in effective bandwidth is possible by using burst rather than element-wise accesses, provided the data is contiguous in memory. We have proposed a memory allocation technique, and provide a proof-of-concept source-to-source compiler pass, that enables such burst transfers by modifying the data layout in external memory. We assess how this technique pushes up the memory throughput, leaving room for exploiting additional parallelism, for a minimal logic overhead. The proposed approach makes it possible to reach 95% of the peak memory bandwidth on a Zynq SoC platform for several representative kernels (iterative stencils, matrix product, convolutions, etc). Our results have been submitted to IEEE Transactions on Computer Aided Design and our submission is under review.

6.2 High-Level Synthesis of Speculative Hardware Accelerators

Participants Steven Derrien, Simon Rokicki, Jean-Michel Gorius.

High Level Synthesis techniques, which compiles C/C++ code directly to hardware circuits, has continuously improved over the last decades. For example, several recent research results have shown how High-Level-Synthesis could be extended to synthesize efficient speculative hardware structures [5]. In particular, speculative loop pipelining appears as a promising approach as it can handle both control-flow and memory speculations within a classical HLS framework. Our last contribution in this topic consists in proposing a fully automated hardware synthesis flow based on a source-to-source compiler that identifies and explores intricate speculation configurations to generate speculative hardware accelerators. We demonstrate that the proposed tool is capable of generating efficient accelerators for several real-life applications, which greatly benefit from the use of speculation. Some of our early results for this work have been submitted to IEEE Micro journal, and is currently under peer reviewing.

6.3 Design Space Exploration for IoT Processors Platforms

Participants Steven Derrien, Simon Rokicki, Jean-Michel Gorius.

The Internet of Things opens many opportunities for new digital products and applications. It also raises many challenges for computer designers: devices are expected to handle larger/bigger computational workloads (e.g., AI-based) while enforcing stringent cost and energy efficiency. The vast majority of IoT platforms rely on low-power Micro-Controller Units families (e.g., ARM Cortex. These MCUs support a same Instruction Set Architecture (ISA) but expose different energy/performance trade-offs thanks to distinct micro-architectures (e.g., the M0 to M7 range in the cortex family). Most existing MCUs rely on proprietary ISAs which prevent third parties to freely implement their own customized micro-architecture and/or deviate from a standardized ISA, therefore hindering innovation. The RISC-V initiative is an effort to address this issue by developing and promoting an open instruction set architecture. The RISC-V ecosystem is quickly growing and has gained a lot of traction for IoT platforms designers, as it permits free customization of both the ISA and the micro-architecture. The problem of customizing/retargeting compilers to a new instruction (or instructions set extension) had been widely studied in the late 90s, and modern compiler infrastructures such as LLVM now offer many facilities for this purpose. However, the problem of automatically synthesizing customized micro-architectures has received much less attention. Although there exist several commercial tools for this purpose, they are based on low-level structural models of the underlying processor pipeline and are not fundamentally different from HDL based approaches (e.g., the processor datapath pipeline organization must be explicit, and hazard management logic is still left to the designer). We are currently looking at novel techniques to bridge the remaining gap between Instruction Set Processor design flows and High-Level-Synthesis tools. More specifically, we aim at taking advantage of speculative loop pipelining to automatically synthesize in order pipelined micro-architectures directly from an Instruction Set Simulator model in C. Although the work is expected to focus on the open-source RISC-V instruction sets, we expect to be able to generalize the approach to more specialized ISA (for cryptographic primitives, packet parsing, etc.).

6.4 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 [26], 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.

6.5 Recursive Polyhedral Equations

Participants Patrice Quinton, Tomofumi Yuki.

Polyhedral equations allow parallel program to be expressed, analyzed, and compiled automatically, but they cannot express divide-and-conquer approaches. This limitation is basically due to the affine nature of the dependence functions imposed by the model. In this research, we addressed how this limitation can be overcome by extending a structured polyhedral equational language to recursive calls of polyhedral programs. Doing so, we preserve the affine property inside a given call, whereas the non-affine part is carried by the recursive expression of subsystem calls. We described the basic mechanisms of this extension, showed that the fundamental results of polyhedral equations hold, in particular, the schedule of such a system can be found automatically. We illustrated this approach on several well known algorithms, including the FFT [36].

6.6 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 [23], 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.

6.7 Training Deep Neural Networks with Low-Precision Accelerators

Participants Silviu Filip, Olivier Sentieys.

The computational workloads associated with training and using Deep Neural Networks (DNNs) pose significant problems from both an energy and an environmental point of view. Designing state-of-the-art neural networks with current hardware can be a several month long process with a significant carbon footprint, equivalent to the emissions of dozens of cars during their lifetimes. If the full potential that deep learning (DL) promises to offer is to be realized, it is imperative to improve existing network training methodologies and the hardware being used by targeting energy efficiency with orders of magnitude reduction. This is equally important for learning on cloud datacenters as it is for learning on edge devices because of communication efficiency and privacy issues. We address this problem at the arithmetic, architecture, and algorithmic levels and explore new mixed numerical precision hardware architectures that are more efficient, both in terms of speed and energy.

In this work, we explore the impact of low-precision operators on training accuracy and overhead, by concurrently developing two frameworks: a software-only GPU-accelerated DNN training system dealing with multi-precision, MPTorch, and a hybrid FPGA-accelerated DNN training system, MPArchimedes, which is implemented with Xilinx Zynq UltraScale+ devices and Vivado HLS [38]. We modified the general matrix multiply (GEMM) layer to train CNN and MLP networks using accelerated low-precision operators. With this change, we have found that existing tools such as QPyTorch produce optimistic accuracy results that are higher than what would be produced by a low-precision edge device. We plan to add support for more accelerated layers and other types of low-precision arithmetic such as posits, logarithmic number systems, or block floating-point, ultimately building a system to explore best practices for mixed-precision DNN training with the lowest area, energy and storage overhead. This work is conducted in collaboration with University of British Columbia, Vancouver, Canada.

We also published a book chapter that explores and reviews how Approximate Computing can improve the performance and energy efficiency of hardware accelerators in Deep Learning applications during inference and training [40].

6.8 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 [29], 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.

We also published a book chapter that reviews low-precision arithmetic operators and custom number representations [41]. This chapter is part of a book on Approximate Computing Techniques [39], Olivier Sentieys from Taran being one of the editors of this book.

6.9 Towards an Arithmetic-Centered Global Approach for Filter Design

Participants Silviu Filip.

Linear time invariant (LTI) digital filters are essential components of modern technology, being used in many applications, ranging from medical equipment and scientific instruments to radar and navigation systems. Depending on the context, they can be implemented either in software, or in hardware when performance and/or power efficiency are critical. In [7], we investigate how the classic filter design and implementation problem in a hardware context can be cast as a single optimization problem and review the current status of the availability and practical use of methods able to solve this problem. We also frame the required advances (in terms of methodology and tools) required to handle the design and synthesis of a wide range of LTI digital filters in practice.

6.10 Fault-Tolerant Microarchitectures

Participants Angeliki Kritikakou, Olivier Sentieys.

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 [30], 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. The approach highlights that a significant number of Multiple-Event-Upsets (MEUs) are derived by faults (SETs) in the combinational logic. These MEUs can be significantly large in size and they not disturb only adjacent bits. Thus, radiation- induced faults should not be modeled only with SEU, but also with (significantly large) MEUs. Our case-study is a RISC-V processor. Obtained results show a more than 8% reduction in masked errors, increasing system failures by more than 55% compared to standard fault injection approaches, which fully validates the hypothesis on the impact of MEUs to the vulnerability and our analysis flow.

6.11 Fault-Tolerant Networks-on-Chip

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

Network-on-Chip has become the main interconnect in the multicore/manycore era since the beginning of this decade. However, these systems become more sensitive to faults due to transistor shrinking size. In parallel, approximate computing appears as a new computation model for applications since several years. The main characteristic of these applications is to support the approximation of data, both for computations and for communications. To exploit this specific application property, we develop a fault-tolerant NoC to reduce the impact of faults on the data communications. To address this problem, we consider multiple permanent faults on router which cannot be managed by Error-Correcting Codes (ECCs), 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 Least Significant Bits (LSBs) instead of MSBs reducing the errors impact. In [22], 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 Mean Square Error (MSE)of the centroid position in the K-means clustering algorithm with a limited area cost and power consumption. To decrease hardware costs, we proposed a region-based bit-shuffling technique in [31], applied at a coarse-grain level, that trades off fault mitigation efficiency in order to save hardware

costs. The obtained results show that the area and power overheads can be reduced from 48% to 33% and from 34% to 22%, respectively, with a small impact on the MSE.

6.12 Fault-Tolerant Task Deployment onto Multicore Systems

Participants Emmanuel Casseau, Minyu Cui, Angeliki Kritikakou.

Task deployment plays an important role in the overall system performance, especially for complex architectures, since it affects not only the energy consumption but also the real-time response and reliability of the system. We are focusing on how to map and schedule tasks onto homogeneous processors under faults at design time. Dynamic Voltage/Frequency Scaling (DVFS) is typically used for energy saving, but with a negative impact on reliability, especially when the frequency is low. Using high frequencies to meet reliability and real-time constraints leads to high energy consumption, while multiple replicas at lower frequencies may increase energy consumption. To minimize energy consumption, enhancing reliability, without violating real-time constraints, we propose an approach that combines distinct reliability enhancement techniques, under task-level, processor-level and system-level DVFS. The problem is initially formulated as Integer Non-Linear Programming and equivalently transformed to a Mixed Integer Linear Programming problem to be optimally solved [27]. To cope with the complexity of such problem, we are currently working on mapping heuristics in order to reduce the time required to find a solution and thus enhance the scalability of the proposed approach.

Furthermore, in [32] a task deployment approach is proposed for multicore architectures with homogeneous cores connected with Network-on-Chip (NoC). The goal is to optimize the overall system energy consumption, including computation of the cores and communication of the NoC, under task reliability and real-time constraints. More precisely, the task deployment approach combines task allocation and scheduling, frequency assignment, task duplication, and multipath data routing. The task deployment problem is formulated using mixed-integer non-linear programming. To find the optimal solution, the original problem is equivalently transformed to mixed-integer linear programming, and solved by state-of-the-art solvers. Furthermore, a decomposition-based heuristic, with low computational complexity, is proposed to deal with scalability. This work is done in collaboration with Lei Mo School of Automation, Southeast University (China).

6.13 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 [24], 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.

6.14 Optical Network-on-Chip for error resilient applications

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

The energy consumption of manycore is dominated by data transfers, which calls for energy- efficient and high-bandwidth interconnects. Classical electrical NoC solutions suffer from low scalability and low performance when the number of cores to connect becomes high. To tackle this challenge, integrated optics appears as promising technology to overcome the bandwidth limitations of electrical interconnects. However, this technology suffers from high power overhead related to low efficiency lasers. From these observations, the concept of approximate communications appears as interesting technique to reduce the power of lasers. In this context, 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 too 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. Furthermore, we also define a distance aware technique which takes account of both the communication distance and the quality of service to compute the laser power [21]. From these contributions, we have developed a simulation platform, based on Sniper, and we show that our solution is scalable and leads to 10% reduction in the total energy consumption, 35× reduction in the laser driver size, and 10× reduction in the laser controller compared to state-of-the-art solutions.

6.15 Dynamic Optical Network-on-Chip based Phase Change Material

Participants Joel Ortiz Sosa, Cédric Killian.

A key challenge for the deployment of nanophotonic interconnects is their high static power, which is induced by signal losses and devices calibration. To tackle this challenge, we propose to use Phase Change Material (PCM) to configure optical paths between writers and readers. The non-volatility of PCM elements and the high contrast between crystalline and amorphous phase states allow to bypass unused readers, thus reducing losses and calibration requirements. We evaluate the efficiency of the proposed PCM-based interconnects using system level simulations carried out with SNIPER manycore simulator. For this purpose, we have modified the simulator to partition clusters according to executed applications. Simulation results show that bypassing readers using PCM leads up to 52% communication power saving [35].

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

Contract with Orange Labs on hardware acceleration on reconfigurable FPGA architectures for next-generation edge/cloud infrastructures. The work program includes: (i) the evaluation of High-Level Synthesis (HLS) tools and the quality of synthesized hardware accelerators, and (ii) time and space sharing of hardware accelerators, going beyond coarse-grained device level allocation in virtualized infrastructures. The two topics are driven from requirements from 5G use cases including 5G LDPC and deep learning LSTM networks for network management.

7.2 Bilateral Grants with Industry

Safran is funding a PhD to study the FPGA implementation of deep convolutional neural network under SWAP (Size, Weight And Power) constraints for detection, classification, image quality improvement of observation systems, and awareness functions (trajectory guarantee, geolocation by cross view alignment) applied to autonomous vehicle. This thesis in particular considers pruning and reduced precision.

Nokia Bell Labs is funding a PhD on FPGA acceleration in the cloud. The goal is to accelerate relational data processing, typically SQL query processing, by leveraging remote memory and remote direct memory access to reduce cloud database services' latency.

7.3 Informal Collaborations with Industry

TARAN collaborates with Mitsubishi Electric R&D Centre Europe (MERCE) on the design and formal verification of Floating-Point Units (FPU).

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 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) and edge devices.

8.1.2 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.

LRS

Title: Loop unRolling Stones: compiling in the polyhedral model

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

Inria contact: Steven Derrien

This collaboration led to two International jointly supervised PhDs (or 'cotutelles' in French) that started in Oct. 2019, one in France (C. Ferry) and one in US (L. Narmour).

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 in the Bachelor and Master ICT degrees.

8.2 International research visitors

8.2.1 Visits of international scientists

Louis Narmour from Colorado State University (CSU) will visit TARAN from Jan. 2022 for two years, in the context of his international jointly supervised PhD (or 'cotutelle' in French) between CSU and Univ. Rennes.

8.2.2 Visits to international teams

Corentin Ferry is visiting Colorado State University (CSU) since Sep. 2021 for one year, in the context of his international jointly supervised PhD (or 'cotutelle' in French) between CSU and Univ. Rennes.

8.3 National initiatives

8.3.1 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: TARAN
- Other partners: INL, 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.

8.3.2 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, TARAN, 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.

8.3.3 ANR Opticall2

Participants Olivier Sentieys, Cédric Killian, Daniel Chillet.

- Program: ANR PRCE
- Project acronym: Opticall2
- Project title: on-chip OPTIcal interconnect for ALL to ALL communications
- Duration: Dec. 2018 - June. 2023
- Coordinator: INL
- Other partners: INL, TARAN, C2N, CEA-LETI, Kalray

The aim of Opticall2 is to design broadcast-enabled optical communication links in manycore architectures at wavelengths around 1.3um. 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 Opticall2 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.

8.3.4 ANR SHNOC

Participants Cédric Killian, Daniel Chillet, Olivier Sentieys, Emmanuel Casseau, Ibrahim Krayem, Yash Aggrawal.

- Program: ANR JCJC (young researcher)
- Project acronym: SHNOC
- Project title: Scalable Hybrid Network-on-Chip
- Duration: Feb. 2019 - Apr. 2024
- P.I.: C. Killian, TARAN

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/sqcm) 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 propose a fast exploration methodology based on analytical models of the hybrid NoC instead of using time consuming manycore simulators. This will allow exploration to 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.

8.3.5 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. Partners: Temento Systems, ONERA, TARAN. For more details see flodam.fr

8.3.6 ANR FASY

Participants Angeliki Kritikakou, Olivier Sentieys.

- Program: ANR JCJC (young researcher)
- Project acronym: FASY
- Project title: FAult-aware timing behaviour for safety-critical multicore SYstems
- Duration: Jan. 2022 - Dec. 2025
- P.I.: K. Kritikakou, TARAN

The safety-critical embedded industries, such as avionics, automobile, robotics and health-care, require guarantees for hard real-time, correct application execution, and architectures with multiple processing elements. While multicore architectures can meet the demands of best-effort systems, the same cannot be stated for critical systems, due to hard-to-predict timing behaviour and susceptibility to reliability threats. Existing approaches design systems to deal with the impact of faults regarding functional behaviors. FASY extends the SoA by answering the two-fold challenge of time-predictable and reliable multicore systems through functional and timing analysis of applications behaviour, fault-aware WCET estimation and design of cores with time-predictable execution, under faults.

8.3.7 ANR Re-Trusting

Participants Olivier Sentieys, Angeliki Kritikakou, Silviu-Ioan Filip.

- Program: ANR PRC
- Project acronym: Re-Trusting

- Project title: REliable hardware for TRUSTworthy artificial INtelliGence
- Duration: Oct. 2021 - Sep. 2025
- Coordinator: INL
- Other partners: LIP6, TARAN, THALES

To be able to run Artificial Intelligence (AI) algorithms efficiently, customized hardware platforms for AI (HW-AI) are required. Reliability of hardware becomes mandatory for achieving trustworthy AI in safety-critical and mission-critical applications, such as robotics, smart healthcare, and autonomous driving. The RE-TRUSTING project develops fault models and performs failure analysis of HW-AIs to study their vulnerability with the goal of "explaining" HW-AI. Explaining HW-AI means ensuring that the hardware is error-free and that the AI hardware does not compromise the AI prediction accuracy and does not bias AI decision-making. In this regard, the project aims at providing confidence and trust in decision-making based on AI by explaining the hardware wherein AI algorithms are being executed.

8.3.8 DGA/INRIA Sniffer

Participants Olivier Sentieys.

- Program: DGA/INRIA joint call on AI
- Project acronym: Sniffer
- Project title: Non-intrusive monitoring of mains operated equipment
- Duration: Feb. 2020 - Mar. 2022
- Partners: TARAN, DGA-MI

Based on the SmartSense platform and on high-frequency traces of the power consumption of individual electrical appliances and building-level power monitoring, the aim of Sniffer is the detection and surveillance of equipment connected to the mains supply.

8.3.9 Labex CominLabs - LeanAI (2021-2024)

Participants Silviu-Ioan Filip (PI), Olivier Sentieys, Steven Derrien.

Recent developments in deep learning (DL) are putting a lot of pressure on and pushing the demand for intelligent edge devices capable of on-site learning. The realization of such systems is, however, a massive challenge due to the limited resources available in an embedded context and the massive training costs for state-of-the-art deep neural networks. In order to realize the full potential of deep learning, it is imperative to improve existing network training methodologies and the hardware being used. LeanAI will attack these problems at the arithmetic and algorithmic levels and explore the design of new mixed numerical precision hardware architectures that are at the same time more energy-efficient and offer increased performance in a resource-restricted environment. The expected outcome of the project includes new mixed-precision algorithms for neural network training, together with open-source tools for hardware and software training acceleration at the arithmetic level on edge devices. Partners: TARAN, LS2N/OGRE, INRIA-LIP/DANTE.

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

General chair, scientific chair

- D. Chillet was the General Chair of ARC'21.
- D. Chillet was the General Chair of HiPEAC Rapido'21.

Member of the organizing committees

- O. Sentieys was the Local Chair of the Organizing Committee of ARC'21.
- A. Kritikakou was the Publicity and Web Chair of ARC'21.
- A. Kritikakou was the Artifact evaluation Co-Chair of ECRTS'21 conference.
- C. Killian was in the Organizing Committee of ARC'21.

9.1.2 Scientific events: selection

Chair of conference program committees

- O. Sentieys was the Chair of the D9 Track on Architectural and Microarchitectural Design at IEEE/ACM DATE 2021.
- O. Sentieys served as a committee member in the IEEE EDAA Outstanding Dissertations Award (ODA).
- S. Derrien was the Program Chair of ARC'21.

Member of the conference program committees

- E. Casseau was a member of the technical program committee of Euro-Par, FPT.
- D. Chillet was member of the technical program committee of HiPEAC Rapido, HiPEAC WRC, DSD, ComPAS, DASIP, ARC.
- S. Derrien was a member of technical program committee of IEEE FPL, IEEE ASAP, IEEE/ACM PACT, ARC.
- A. Kritikakou was a member of technical program committee of IEEE RTSS, IEEE RTAS, ECRTS, SAMOS, HPPC, IEEE/ACM DATE, ARC, CPSCom, ComPAS.
- O. Sentieys was a member of technical program committee of IEEE/ACM DATE, IEEE FPL, ACM ENSSys, ACM SBCCI, IEEE ReConFig, FDL, ARC.
- C. Killian was a member of technical program committee of ACM NOCS.
- S. Rokicki was a member of technical program committee of CASES

9.1.3 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 is an editor for a Special Issue in Elsevier Journal of Parallel and Distributed Computing (JPD) "Parallel and Distributed Computing for Cyber-Physical Systems", MDPI Electronics "Dependability of Emerging Computing Paradigms and Technologies in IoT-oriented Circuits, Architectures and Algorithms", MDPI Sensors "Sensor Facilitated Cyber-Physical Systems".

9.1.4 Invited talks

- O. Sentieys gave an invited talk at HiPEAC Computing Systems Week (CSW), Lyon, France, Oct 1, 2021 on "Approximate Deep Learning Accelerators: Improving performance and energy efficiency of deep-learning hardware accelerators with controlled arithmetic approximations"[37].

- O. Sentieys gave an invited talk at Dagstuhl Seminar 21302 - Approximate Systems on "An Optimization Playground for Precision and Number Representation Tuning"[45].

- O. Sentieys gave an invited talk at IIT Goa, India on "An Optimization Playground for Precision and Number Representation Tuning". This talk is supported by IntelliVIS, a joint research team of IIT Goa and INRIA, Rennes supported by CEFIPRA and DST, India and INRIA Associate Team, France.

- O. Sentieys gave an invited talk at "*Séminaire sur la Tolérance Aux Fautes des Equipements Electroniques pour la Defense*" (STAFEED) on "vulnerability analysis of embedded digital systems: from physics to micro-architecture".

- S. Derrien gave an invited talk at "*Colloque GdR SoC2*" on "Improving performance and energy efficiency of CNN accelerators through overclocking.".

- S. Derrien gave an invited talk at "*Journée Thématique GdR SoC2 : Outils pour la Synthèse de Haut Niveau*" on "Toward Speculative Loop Pipelining for High-Level Synthesis.".

- A. Kritikakou gave an invited talk at "*International workshop sCalable and PrecIse Timing AnaLysis for multiocre platforms (CAPITAL)*" on "Run-time adaptation of task execution in time-critical systems: Challenges and Solutions".

- C. Killian gave an invited talk at "*French symposium on photonic (Optique'21)*" on "Tolerating errors in on-chip nanophotonic interconnects for improved energy efficiency".

9.1.5 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.

9.1.6 Scientific expertise

- O. Sentieys was a member of the ANR Scientific Evaluation Committee CE25 "Software science and engineering - Multi-purpose communication networks, high-performance infrastructure".

9.1.7 Research administration

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

9.2 Teaching - Supervision

9.2.1 Teaching Responsibilities

- E. Casseau is in charge of the Department of “Digital Systems” at ENSSAT Engineering Graduate School.
- D. Chillet is associate director of studies at ENSSAT Engineering Graduate School.
- D. Chillet is the responsible of the ”Embedded Systems” major of the SISEA Master by Research.
- C. Killian is the responsible of the second year of the ”Instrumentation” DUT at IUT, Lannion.
- S. Rokicki is the responsible of the second year in the computer science department of ENS Rennes

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.

9.2.2 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: 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, 48h, ESIR (L3)
- F. Charot: software hardware interfaces, 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, 72h, IUT Lannion (L1)
- C. Killian: automated measurements, 53h, IUT Lannion (L2)
- C. Killian: computer architecture, 6h, IUT Lannion (L3)
- C. Killian: embedded systems, 46h, IUT Lannion (L2)
- C. Killian: microcontrollers, 49h, IUT Lannion (L2)
- A. Kritikakou: principles of computer design, 32h, ISTIC (L3)

- A. Kritikakou: software hardware interfaces, 44h, ISTIC (L3)
- A. Kritikakou: C and unix programming languages, 76h, ISTIC (L3)
- A. Kritikakou: operating systems, 48h, ISTIC (L3)
- O. Sentieys: VLSI integrated circuit design, 24h, ENSSAT (M1)
- O. Sentieys: VHDL and logic synthesis, 18h, ENSSAT (M1)
- S. Rokicki: C Programming, 24h, ENS Rennes

9.2.3 PhD Supervision

- PhD: Romain Mercier, Multiple Fault Mitigation in Network-on-Chip Architectures Through a Bit-Shuffling Method, Dec. 2021, D. Chillet, C. Killian, A. Kritikakou.
- 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: Sami Ben Ali, Efficient Low-Precision Training for Deep Learning Accelerators, Jan. 2022, O. Sentieys.
- 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 Univ Rennes 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: Jean-Michel Gorius, Speculative Software Pipeline for Micro-Architecture Synthesis, Oct. 2021, S. Derrien, S. Rokicki.
- 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: Seungah Lee, Efficient Designs of On-Board Heterogeneous Embedded Systems for Space Applications, Nov. 2021, A. Kritikakou, E. Casseau, R. Salvador (Centrale-Supelec), O. Sentieys.
- PhD in progress: Amélie Marotta, Emp-error: EMFI-Resilient RISC-V Processor, Oct. 2021, O. Sentieys, R. Lashermes (LHS), Rachid Dafali (DGA).
- PhD in progress: Thibaut Marty, Compiler support for speculative custom hardware accelerators, Sep. 2017, T. Yuki, S. Derrien.
- PhD in progress: Louis Narmour, Revisiting memory allocation in the 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.

10 Scientific production

10.1 Major publications

[1] B. Barrois and O. Sentieys. 'Customizing Fixed-Point and Floating-Point Arithmetic - A Case Study in K-Means Clustering'. In: SiPS 2017 - IEEE International Workshop on Signal Processing Systems. Lorient, France, Oct. 2017. URL: https://hal.inria.fr/hal-01633723.

[2] B. Barrois, O. Sentieys and D. Ménard. 'The Hidden Cost of Functional Approximation Against Careful Data Sizing – A Case Study'. In: Design, Automation & Test in Europe Conference & Exhibition (DATE 2017). Lausanne, Switzerland, 2017. DOI: 10.23919/date.2017.7926979. URL: https://hal.inria.fr/hal-01423147.

[3] N. Brisebarre, G. Constantinides, M. Ercegovac, S.-I. Filip, M. Istoan and J.-M. Muller. 'A High Throughput Polynomial and Rational Function Approximations Evaluator'. In: ARITH 2018 - 25th IEEE Symposium on Computer Arithmetic. Amherst, MA, United States: IEEE, 25th June 2018, pp. 99–106. DOI: 10.1109/ARITH.2018.8464778. URL: https://hal.inria.fr/hal-01774364.

[4] G. Deest, T. Yuki, S. Rajopadhye and S. Derrien. 'One size does not fit all: Implementation trade-offs for iterative stencil computations on FPGAs'. In: FPL - 27th International Conference on Field Programmable Logic and Applications. Gand, Belgium: IEEE, 4th Sept. 2017. DOI: 10.23919/FPL.2017.8056781. URL: https://hal.inria.fr/hal-01655590.

[5] 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), pp. 4229–4239. DOI: 10.1109/TCAD.2020.3012866. URL: https://hal.archives-ouvertes.fr/hal-02949516.

[6] S. Derrien, S. Rajopadhye, P. Quinton and T. Risset. 'High-Level Synthesis of Loops Using the Polyhedral Model'. In: *High-Level Synthesis : From Algorithm to Digital Circuit*. Spinger, 2008, pp. 215–230. URL: https://hal.archives-ouvertes.fr/hal-00410719.

[7] F. de Dinechin, S.-I. Filip, L. Forget and M. Kumm. 'Table-Based versus Shift-And-Add constant multipliers for FPGAs'. In: ARITH 2019 - 26th IEEE Symposium on Computer Arithmetic. Kyoto, Japan: IEEE, 10th June 2019, pp. 1–8. URL: https://hal.inria.fr/hal-02147078.

[8] A. Floch, T. Yuki, A. El-Moussawi, A. Morvan, K. Martin, M. Naullet, M. Alle, L. L'Hours, N. Simon, S. Derrien, F. Charot, C. Wolinski and O. Sentieys. 'GeCoS: A framework for prototyping custom hardware design flows'. In: 13th IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM). Eindhoven, Netherlands: IEEE, 23rd Sept. 2013, pp. 100–105. DOI: 10.1109/SCAM.2013.6648190. URL: https://hal.inria.fr/hal-00921370.

[9] M. Fyrbiak, S. Rokicki, N. Bissantz, R. Tessier and C. Paar. 'Hybrid Obfuscation to Protect against Disclosure Attacks on Embedded Microprocessors'. In: *IEEE Transactions on Computers* (2017). URL: https://hal.inria.fr/hal-01426565.

[10] M. Gueguen, O. Sentieys and A. Termier. 'Accelerating Itemset Sampling using Satisfiability Constraints on FPGA'. In: DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe. Florence, Italy: IEEE, 25th Mar. 2019, pp. 1046–1051. DOI: 10.23919/DATE.2019.8714932. URL: https://hal.inria.fr/hal-01941862.

[11] 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: IEEE, 9th Mar. 2020, pp. 1–6. URL: https://hal.inria.fr/hal-02387232.

[12] 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.

[13] J. Luo, C. Killian, S. Le Beux, D. Chillet, O. Sentieys and I. O'Connor. 'Offline Optimization of Wavelength Allocation and Laser Power in Nanophotonic Interconnects'. In: *ACM Journal on Emerging Technologies in Computing Systems* 14.2 (27th July 2018), pp. 1–19. DOI: 10.1145/3178453. URL: https://hal.inria.fr/hal-01934870.

[14] 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* 39.12 (Mar. 2020), pp. 4777–4790. DOI: 10.1109/TCAD.2020.2981056. URL: https://hal.inria.fr/hal-03094811.

[15] D. Ménard, G. Caffarena, J. A. Lopez, D. Novo and O. Sentieys. 'Analysis of Finite Word-Length Effects in Fixed-Point Systems'. In: *Handbook of Signal Processing Systems*. 2019, pp. 1063–1101. DOI: 10.1007/978-3-319-91734-4_29. URL: https://hal.inria.fr/hal-01941888.

[16] 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: IEEE, 6th July 2020, pp. 328–333. DOI: 10.1109/ISVLSI49217.2020.00067. URL: https://hal.archives-ouvertes.fr/hal-02927455.

[17] R. Psiakis, A. Kritikakou and O. Sentieys. 'Fine-Grained Hardware Mitigation for Multiple Long-Duration Transients on VLIW Function Units'. In: DATE 2019 - 22nd IEEE/ACM Design, Automation and Test in Europe. Florence, Italy: IEEE, 25th Mar. 2019, pp. 976–979. DOI: 10.23919/DATE.2019.8714899. URL: https://hal.inria.fr/hal-01941860.

[18] 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: IEEE, 9th Mar. 2020, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-02396631.

[19] S. Rokicki, E. Rohou and S. Derrien. 'Hybrid-DBT: Hardware/Software Dynamic Binary Translation Targeting VLIW'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (8th Aug. 2018), pp. 1–14. DOI: 10.1109/TCAD.2018.2864288. URL: https://hal.archives-ouvertes.fr/hal-01856163.

10.2 Publications of the year

International journals

[20] P. Dobiáš, E. Casseau and O. Sinnen. 'Improving the CubeSat Reliability Thanks to a Multiprocessor System using Fault Tolerant Online Scheduling'. In: *Microprocessors and Microsystems: Embedded Hardware Design (MICPRO)* 85 (Sept. 2021), pp. 1–12. DOI: 10.1016/j.micpro.2021.104312. URL: https://hal.inria.fr/hal-03317768.

[21] J. Lee, C. Killian, S. Le Beux and D. Chillet. 'Distance-aware Approximate Nanophotonic Interconnect'. In: *ACM Transactions on Design Automation of Electronic Systems* 27.2 (31st Mar. 2022), pp. 1–30. DOI: 10.1145/3484309. URL: https://hal.inria.fr/hal-03500153.

[22] R. Mercier, C. Killian, A. Kritikakou, Y. Helen and D. Chillet. 'BiSuT: A NoC-Based Bit-Shuffling Technique for Multiple Permanent Faults Mitigation'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (30th July 2021), pp. 1–14. DOI: 10.1109/TCAD.2021.3101406. URL: https://hal.inria.fr/hal-03379489.

[23] 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* 8.10 (May 2021), pp. 8246–8258. DOI: 10.1109/JIOT.2020.3044910. URL: https://hal.archives-ouvertes.fr/hal-03103821.

[24] 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* 40.8 (2021), pp. 1559–1572. DOI: 10.1109/TCAD.2020.3025063. URL: https://hal.inria.fr/hal-03119369.

International peer-reviewed conferences

[25] E. Casseau, P. Dobias, O. Sinnen, G. Rodrigues, F. Kastensmidt, A. Savino, S. Di Carlo, M. Rebaudengo and A. Bosio. 'Special Session: Operating Systems under test: an overview of the significance of the operating system in the resiliency of the computing continuum'. In: VTS 2021 - 39th IEEE VLSI Test Symposium. San Diego, United States: IEEE, 25th Apr. 2021, pp. 1–10. DOI: 10.1109/VTS50974.2021.9441042. URL: https://hal.archives-ouvertes.fr/hal-03266808.

[26] T. Cong and F. Charot. 'Design Space Exploration of Heterogeneous-Accelerator SoCs with Hyperparameter Optimization'. In: ASP-DAC 2021 - 26th Asia and South Pacific Design Automation Conference. Virtual Conference, Japan, 18th Jan. 2021, pp. 1–6. URL: https://hal.inria.fr/hal-03119732.

[27] M. Cui, A. Kritikakou, L. Mo and E. Casseau. 'Fault-Tolerant Mapping of Real-Time Parallel Applications under multiple DVFS schemes'. In: RTAS 2021 - 27th IEEE Real-Time and Embedded Technology and Applications Symposium. Samos Island, Greece: IEEE, 18th May 2021, pp. 387–399. DOI: 10.1109/RTAS52030.2021.00038. URL: https://hal.inria.fr/hal-03501313.

[28] F. de Dinechin, S.-I. Filip, M. Kumm and A. Volkova. 'Towards Arithmetic-Centered Filter Design'. In: ARITH 2021 - 28th IEEE Symposium on Computer Arithmetic. Torino, Italy, 14th June 2021, pp. 1–4. URL: https://hal.inria.fr/hal-03220258.

[29] 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: IEEE, 1st Feb. 2021, pp. 1–6. URL: https://hal.inria.fr/hal-03119548.

[30] A. Kritikakou, O. Sentieys, G. Hubert, Y. Helen, J.-F. Coulon and P. Deroux-Dauphin. 'FLODAM: Cross-Layer Reliability Analysis Flow for Complex Hardware Designs'. In: 25th IEEE/ACM Design, Automation and Test in Europe (DATE). Antwerp, Belgium, Mar. 2022, pp. 1–6. URL: https://hal.archives-ouvertes.fr/hal-03485386.

[31] R. Mercier, C. Killian, A. Kritikakou, Y. Helen and D. Chillet. 'A Region-Based Bit-Shuffling Approach Trading Hardware Cost and Fault Mitigation Efficiency'. In: DFT 2021 - 34th IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems. athens, Greece: IEEE, 6th Oct. 2021, pp. 1–4. DOI: 10.1109/DFT52944.2021.9568366. URL: https://hal.archives-ouvertes.fr/hal-03500395.

[32] L. Mo, Q. Zhou, A. Kritikakou and J. Liu. 'Energy Efficient, Real-time and Reliable Task Deployment on NoC-based Multicores with DVFS'. In: IEEE/ACM Design, Automation and Test in Europe (DATE). IEEE/ACM Design, Automation and Test in Europe (DATE). Antwerp, Belgium, 14th Mar. 2022. URL: https://hal.archives-ouvertes.fr/hal-03500332.

[33] V. L. Nguyen Huu, J. Lallet, E. Casseau and L. D'Orazio. 'MASCARA-FPGA cooperation model: Query Trimming through accelerators'. In: SSDBM 2021 - 33rd International Conference on Scientific and Statistical Database Management. Tampa, United States: ACM, 6th July 2021, pp. 203–208. DOI: 10.1145/3468791.3468795. URL: https://hal.inria.fr/hal-03503635.

[34] O. Sentieys, S.-I. Filip, D. Briand, D. Novo, E. Dupuis, I. O'Connor and A. Bosio. 'AdequateDL: Approximating Deep Learning Accelerators'. In: DDECS 2021 - 24th International Symposium on Design and Diagnostics of Electronic Circuits and Systems. Vienna (virtual), Austria: IEEE, 2021, pp. 37–40. DOI: 10.1109/DDECS52668.2021.9417026. URL: https://hal.archives-ouvertes.fr/hal-03266861.

[35] P. Zolfaghari, J. Ortiz, C. Killian and S. Le Beux. 'Non-Volatile Phase Change Material based Nanophotonic Interconnect'. In: IEEE/ACM Design, Automation and Test in Europe (DATE). IEEE/ACM Design, Automation and Test in Europe (DATE) 2022. Antwerp, Belgium, 1st Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03512251.

Conferences without proceedings

[36] P. Quinton and T. Yuki. 'Representing Non-Affine Parallel Algorithms by means of Recursive Polyhedral Equations'. In: IMPACT 2021. Budapest, Hungary, 20th Jan. 2021. URL: https://hal.archives-ouvertes.fr/hal-03518569.

[37] O. Sentieys. 'Approximate Deep Learning Accelerators: Improving performance and energy efficiency of deep-learning hardware accelerators with controlled arithmetic approximations'. In: CSW 2021 - HiPEAC Computing Systems Week. Lyon, France, Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03494932.

[38] M. Tatsumi, Y. Xie, C. White, S.-I. Filip, O. Sentieys and G. Lemieux. 'MPTorch and MPArchimedes: Open Source Frameworks to Explore Custom Mixed- Precision Operations for DNN Training on Edge Devices'. In: ROAD4NN 2021 - 2nd ROAD4NN Workshop: Research Open Automatic Design for Neural Networks. San Francisco, United States, 5th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03494256.

Scientific books

[39] A. Bosio, D. Menard and O. Sentieys. *Approximate Computing Techniques: From Component- to Application-Level*. Springer, 2022. URL: https://hal.archives-ouvertes.fr/hal-03494868.

Scientific book chapters

[40] E. Dupuis, S.-I. Filip, O. Sentieys, D. Novo, I. O'Connor and A. Bosio. 'Approximations in Deep Learning'. In: *Approximate Computing Techniques - From Component- to Application-Level*. 2022. URL: https://hal.archives-ouvertes.fr/hal-03494874.

[41] O. Sentieys and D. Menard. 'Customizing Number Representation and Precision'. In: *Approximate Computing Techniques - From Component- to Application-Level*. Springer, 2022. URL: https://hal.archives-ouvertes.fr/hal-03494872.

Doctoral dissertations and habilitation theses

[42] R. Mercier. 'Multiple Fault Mitigation in Network-on-Chip Architectures Through A Bit-Shuffling Method'. Université de Rennes 1, 17th Dec. 2021. URL: https://hal.inria.fr/tel-03500147.

Reports & preprints

[43] P. Dobiáš, E. Casseau and O. Sinnen. *Comparison of Enhancing Methods for Primary/Backup Approach Meant for Fault Tolerant Scheduling*. Univ Rennes, Inria, CNRS, IRISA, France, 26th Oct. 2021. URL: https://hal.inria.fr/hal-03405142.

[44] L. Mo, A. Kritikakou and X. Li. *Energy-Efficient, Reliable and QoS-Aware Task Mapping on Cyber-Physical Systems*. IEEE Technical Committee on Cyber-Physical Systems, Aug. 2021. URL: https://hal.inria.fr/hal-03419313.

Other scientific publications

[45] O. Sentieys. *An Optimization Playground for Precision and Number Representation Tuning*. July 2021. URL: https://hal.archives-ouvertes.fr/hal-03491936.

10.3 Cited publications

[46] 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.

[47] 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.

[48] V. Chippa, S. Chakradhar, K. Roy and A. Raghunathan. 'Analysis and characterization of inherent application resilience for approximate computing'. In: *50th ACM/IEEE Design Automation Conf. (DAC)*. May 2013, pp. 1–9.

[49] 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.

[50] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam and D. Burger. 'Dark Silicon and the End of Multicore Scaling'. In: *Proc. 38th Int. Symp. on Computer Architecture (ISCA)*. San Jose, California, USA, 2011, pp. 365–376. DOI: 10.1145/2000064.2000108. URL: http://doi.acm.org/10.1145/2000064.2000108.

[51] R. Hameed et al. 'Understanding Sources of Ineffciency in General-purpose Chips'. In: *Commun. ACM* 54.10 (Oct. 2011), pp. 85–93. DOI: 10.1145/2001269.2001291. URL: http://doi.acm.org/10.1145/2001269.2001291.

[52] E. Ibe et al. 'Impact of Scaling on Neutron-Induced Soft Error in SRAMs From a 250 Nm to a 22 Nm Design Rule'. In: *IEEE Trans. on Elect. Dev.* 57.7 (2010), pp. 1527–1538.

[53] H. Lee, D. Nguyen and J. Lee. 'Optimizing Stream Program Performance on CGRA-based Systems'. In: *52nd IEEE/ACM Design Automation Conference*. 2015, 110:1–110:6.

[54] S. Mittal. 'A survey of techniques for approximate computing'. In: *ACM Computing Surveys (CSUR)* 48.4 (2016), pp. 1–33.

[55] A. Putnam *et al.* 'A reconfigurable fabric for accelerating large-scale datacenter services'. In: *ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. June 2014, pp. 13–24.

[56] S. Rehman et al. *Reliable Software for Unreliable Hardware: A Cross Layer Perspective*. Springer, 2016.

[57] N. Seifert et al. 'Soft Error Susceptibilities of 22 Nm Tri-Gate Devices'. In: *IEEE Trans. on Nuclear Science* 59 (2012), pp. 2666–2673.

[58] V. Sze, Y.-H. Chen, T.-J. Yang and J. S. Emer. 'Efficient processing of deep neural networks: A tutorial and survey'. In: *Proceedings of the IEEE* 105.12 (2017), pp. 2295–2329.

[59] V. Vargas et al. 'Radiation Experiments on a 28 nm Single-Chip Many-Core Processor and SEU Error-Rate Prediction'. In: *IEEE Trans. on Nuclear Science* 64.1 (Jan. 2017), pp. 483–490.