UMR IRISA

Activity Report 2023

Team TARAN

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

Joint team with Centre Inria de l'Université de Rennes

D3 – Architecture

Université de Rennes | ENS rennes | INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES | cnrs | Inria | Inserm | CentraleSupélec | IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom | Université Bretagne Sud ubs

Contents

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

- Olivier Sentieys [Team leader, INRIA, Senior Researcher, from Sep 2023, HDR]
- François Charot [INRIA, Researcher]
- Fernando Fernandes Dos Santos [INRIA, Starting Research Position]
- Silviu-Ioan Filip [INRIA, Researcher]
- Marcello Traiola [INRIA, Researcher]

Faculty Members

- Olivier Sentieys [Team leader, UNIV RENNES, Professor, until Aug 2023, HDR]
- Emmanuel Casseau [UNIV RENNES, Professor, HDR]
- Daniel Chillet [UNIV RENNES, Professor, HDR]
- Steven Derrien [UNIV RENNES, Professor, HDR]
- Cédric Killian [UNIV RENNES, Associate Professor, until Aug 2023, HDR]
- Angeliki Kritikakou [UNIV RENNES, Associate Professor, HDR]
- Bertrand Le Gal [UNIV RENNES, Associate Professor, from Sep 2023, HDR]
- Patrice Quinton [ENS RENNES, Emeritus]
- Simon Rokicki [ENS RENNES, Associate Professor]

Post-Doctoral Fellows

- Abhijit Das [UNIV RENNES, Post-Doctoral Fellow, until Feb 2023]
- Remi Garcia [UNIV RENNES, Post-Doctoral Fellow, from Dec 2023]

PhD Students

- Hamza Amara [UNIV RENNES]
- Herinomena Andrianatrehina [INRIA]
- Gaetan Barret [ORANGE, CIFRE]
- Sami Ben Ali [INRIA]
- Benoit Coqueret [THALES, CIFRE]
- Leo De La Fuente [CEA]
- Paul Estano [INRIA]
- Corentin Ferry [UNIV RENNES]
- Cedric Gernigon [INRIA]
- Jean-Michel Gorius [UNIV RENNES]
- Wilfread Guilleme [INRIA]

- Ibrahim Krayem [UNIV RENNES]
- Seungah Lee [UNIV RENNES]
- Dylan Leothaud [UNIV RENNES, from Sep 2023]
- Guillaume Lomet [INRIA]
- Amélie Marotta [INRIA]
- Louis Narmour [UNIV RENNES]
- Pegdwende Nikiema [UNIV RENNES]
- Léo Pajot [KEYSOM SAS, CIFRE, from Sep 2023]
- Leo Pradels [INRIA, from Dec 2023]
- Leo Pradels [SAFRAN, until Nov 2023]
- Baptiste Rossigneux [CEA]
- Louis Savary [INRIA]

Technical Staff

- Sonia Barrios Pereira [INRIA, Engineer, until Apr 2023]
- Ludovic Claudepierre [INRIA, Engineer, until Oct 2023]
- Romain Mercier [UNIV RENNES, Engineer, until Feb 2023]
- Joseph Paturel [INRIA, Engineer]
- Dikshanya Lashmi Ramaswamy [INRIA, Engineer]
- Etienne Tehrani [INRIA, Engineer, from Dec 2023]

Interns and Apprentices

- Fierte Balbina Aguessy [UNIV RENNES, Intern, from Jul 2023 until Aug 2023]
- Thomas Feuilletin [UNIV RENNES, Intern, from May 2023 until Aug 2023]
- Hugo Groussin [UNIV RENNES, Intern, from Jun 2023 until Sep 2023]
- Saaswath Lakshmanan Nachiappa [UNIV RENNES, Intern, from Jun 2023 until Jul 2023]
- Dylan Leothaud [INRIA, Intern, until Jul 2023]
- Nathan Merillon [ENS Rennes, Intern, from Jun 2023 until Aug 2023]
- Hadrien Moulherat [UNIV RENNES, Intern, from May 2023 until Jul 2023]
- Melanie Romano [UNIV RENNES, Intern, from May 2023 until Aug 2023]
- Lucas Roquet [UNIV RENNES, Intern, from May 2023 until Aug 2023]
- Oleksii Tkachenko [UNIV RENNES, Intern, from Jul 2023 until Aug 2023]

Administrative Assistants

- Emilie Carquin [UNIV RENNES]
- Nadia Derouault [INRIA]

Visiting Scientists

- Pavitra Bhade [IIT Goa, from Feb 2023 until Feb 2023]
- Hendrik Wohrle [FH Dortmund, from Apr 2023 until Jun 2023]
- Jinyi Xu [CSC Scholarship, until Jun 2023]

External Collaborator

- Guillaume Didier [DGA, from Mar 2023]

2 Overall objectives

Energy efficiency has now become one of the main requirements for virtually all computing platforms [62]. We now have an opportunity 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 [61, 60], 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 [56]. 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 run 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, B. Le Gal, P. Quinton, S. Rokicki, 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% [66], and the increasing need for acceleration of deep learning workloads [69].

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 [57, 64]. 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, M. Traiola.

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 [7]. 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 [65]. 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 [58].

3.3 Resilient Computing

Key Investigators: E. Casseau, D. Chillet, F. F. Dos Santos, A. Kritikakou, O. Sentieys, M. Traiola.
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 [63], leading to dramatically reduced yield rates [67]. Transient errors caused by particles or radiations will also more and more often occur during execution [70, 68], 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 [9, 8, 4], 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, O. Sentieys, M. Traiola.
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 [10] and we will continue to explore new application domains where HW acceleration is pertinent.

5 Highlights of the year

5.1 Awards

Angeliki Kritikakou has been appointed Junior Member of the Institut universitaire de France (IUF) by order of the French Ministère de l'Enseignement supérieur et de la Recherche for a period of 5 years, starting from October 1 2023. She holds an Innovation Chair.

Fernando Fernandes dos Santos won the McCluskey doctoral thesis award at IEEE International Test Conference (ITC) 2023 in Anaheim, CA, USA.

Silviu Filip received the Best Paper Award at the IEEE Symposium on Computer Arithmetic 2023 [27] .

Olivier Sentieys received the *Trophée Valorisation du Campus Innovation de l'Université de Rennes* in Digital Science. D. Ménard (INSA Rennes) and J. Bonnot (WeDoLow) are co-recipients of the Award.

6 New software, platforms, open data

6.1 New software

6.1.1 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 targeted at program transformations mainly for High-Level-Synthesis tools. Gecos uses the Eclipse Modeling Framework (EMF) as an underlying infrastructure. Gecos is open-source and is hosted on the Inria gitlab. The Gecos infrastructure is still under very active development and serves as a backbone infrastructure to several research projects of the group.

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.

News of the Year: With the recent work on the Speculative HLS project and the new ANR LOTR, we have extended the tool to integrate some new analysis and transformation based on the Circt project (https://circt.llvm.org). We are also moving toward generating verilog for a subset of input C code. The objective is to be able to generate hardware with a fully open-source toolchain.

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

Publication: hal-03714101

Contact: Steven Derrien

Participants: Simon Rokicki, Dylan Leothaud, Jean-Michel Gorius, Steven Derrien

Partner: Université de Rennes 1

6.1.2 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

Participants: Olivier Sentieys, Guillermo Enrique Andrade Barroso, Mickael Le Gentil, Sonia Barrios Pereira

6.1.3 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

Participants: Olivier Sentieys, Van Phu Ha, Tomofumi Yuki, Ali Hassan El Moussawi

6.2 New platforms

6.2.1 MPTorch: a PyTorch-based framework for simulating custom precision DNN training

Participants: Silviu-Ioan Filip, Paul Estano.

KEYWORDS: Computer architecture, Arithmetic, Custom Floating-point, Deep learning, Multiple-Precision

SCIENTIFIC DESCRIPTION: MPTorch is a wrapper framework built atop PyTorch that is designed to simulate the use of custom/mixed precision arithmetic in PyTorch, especially for DNN training.

FUNCTIONAL DESCRIPTION: MPTorch reimplements the underlying computations of commonly used layers for CNNs (e.g. matrix multiplication and 2D convolutions) using user-specified floating-point formats for each operation (e.g. addition, multiplication). All the operations are internally done using IEEE-754 32-bit floating-point arithmetic, with the results rounded to the specified format.

- Contact: Silviu-Ioan Filip
- URL: mptorch on github

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

Participants: Silviu-Ioan Filip, Matei Istoan.

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.

- Contact: Silviu-Ioan Filip
- Partners: Univ Rennes, Imperial College London
- URL: emethod on github

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

Participants: Silviu-Ioan Filip, Martin Kumm, Anastasia Volkova.

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).

- Contact: Silviu-Ioan Filip
- Partners: Univ Rennes, Université de Nantes, Fulda University of Applied Sciences
- URL: firopt on gitlab

6.2.4 rminimax: a tool for designing machine-efficient rational approximations of mathematical functions

Participants: Silviu-Ioan Filip.

KEYWORDS: Computer Arithmetic, Function Approximation, Rational and Polynomial Functions, Mathematical Libraries

SCIENTIFIC DESCRIPTION: rminimax is a C++ library for designing L^∞-based rational approximations of mathematical function, with both real and machine-representable coefficients (such as IEEE-754 floating-point formats). The output of the tool is intended for use inside custom mathematical function accelerators (both hardware and software-based), for instance in an FPGA context or for mathematical libraries like the `libm` of the C language.

- Contact: Silviu-Ioan Filip
- Partners: Univ Rennes
- URL: rminimax on gitlab

6.2.5 Hybrid-DBT

Participants: Simon Rokicki, Louis Savary, Steven Derrien.

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.

- Contact: Simon Rokicki
- Partners: Univ Rennes
- URL: HybridDBT on github

6.2.6 Comet

Participants: Simon Rokicki, Olivier Sentieys, Joseph Paturel.

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.

- Contact: Simon Rokicki
- Partners: Univ Rennes
- URL: Comet on gitlab

7 New results

7.1 Improving Memory Throughput of Hardware Accelerators

Participants: Steven Derrien, Corentin Ferry.

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 accelerator 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 provided 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.) [15].

7.2 High-Level Synthesis of Speculative Hardware Accelerators

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

High Level Synthesis (HLS) 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 [1]. 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 unified memory speculation framework, which allows aggressive scheduling and high-throughput accelerator synthesis in the presence of complex memory dependencies. In a publication just accepted in CC'24 [31], we show that our technique can generate high-throughput designs for various applications and describe a complete implementation inside the Gecos HLS toolchain.

7.3 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 hardware description languages (HDL) based approaches (e.g., the processor datapath pipeline organization must be explicit, and hazard management logic is still left to the designer).

Our recent work on the subject aims to demonstrate how the available features of HLS can be deployed in designing various pipelined processors micro-architecture. Our approach takes advantage of the capabilities of existing HLS tools and employs multi-threading and dynamic scheduling techniques to overcome the limitation of HLS in pipelining a processor from an Instruction Set Simulator written in C. This work has been published and presented in ARC'23 [34].

The ANR project LOTR is also focused on this topic. The proposed flow operates on a description of the processor Instruction Set Architecture (ISA). It can automatically infer synthesizable Register Transfer Level (RTL) descriptions of a large number of microarchitecture variants with different performance/cost trade-offs. In addition, the flow will integrate two domain-specific toolboxes dedicated to the support of timing predictability (for safety-critical systems) and security (through hardware protection).

7.4 High-Level Synthesis-Based On-board Payload Data Processing considering the Roofline Model

Participants: Seungah Lee, Olivier Sentieys, Angeliki Kritikakou, Emmanuel Casseau.

On-board payload data processing can be performed by developing space-qualified heterogeneous Multiprocessor Systemon-Chips (MPSoCs). In [36], we present key compute-intensive payload algorithms, based on a survey with space science researchers, including the two-dimensional Fast Fourier Transform (2-D FFT). Also, we propose to perform design space exploration by combining the roofline performance model with High-Level Synthesis (HLS) for hardware accelerator architecture design. The roofline model visualizes the limits of a given architecture regarding Input/Output (I/O) bandwidth and computational performance, along with the achieved performance for different implementations. HLS is an interesting option in developing FPGA-based onboard processing applications for payload teams that need to adjust architecture specifications through design reviews and have limited expertise in Hardware Description Languages (HDLs). In this work, we focus on an FPGA-based MPSoC thanks to recently released radiation-hardened heterogeneous embedded platforms.

This work is done in collaboration with Julien Galizzi, CNES (France).

7.5 Energy Constrained and Real-Time Imprecise Computation Tasks Mapping on Multicore Systems

Participants: Olivier Sentieys, Angeliki Kritikakou.

Heterogeneous multicore architectures have become one of the most widely used hardware platforms for embedded systems, where time, energy, and system QoS are the major concerns. The Imprecise Computation (IC) model splits a task into mandatory and optional parts, allowing the trade-off of the above issues. However, existing approaches, to maximize system QoS (Quality-of-Service) under time or energy constraints, use a linear function to model system QoS. Therefore, they become unsuitable for general applications, whose QoS is modeled by a concave function. To deal with this limitation, we addresses in [21] the Mixed-Integer Non-Linear Programming (MINLP) problem of mapping IC tasks to a set of heterogeneous cores by concurrently deciding which processor executes each task and the number of cycles of optional tasks (i.e., task allocation and scheduling), under real-time and energy supply constraints. Furthermore, as existing solution algorithms either demand high time complexity or only achieve feasible solutions, we propose a novel approach based on problem transformation and dual decomposition that finds an optimal solution while avoiding high computational complexity. Simulation

results show that the proposed approach achieves 98% performance of the optimization solver Gurobi, but only with 19.8% of its computation time.

This work is done in collaboration with Lei Mo, School of Automation, Southeast University (China).

7.6 Run-time Mechanisms for Mitigating Mode-Switch in Mixed-Critical Embedded Systems

Participants: Angeliki Kritikakou.

Mixed-critical systems consist of applications with different criticality. In these systems, different confidence levels of Worst-Case Execution Time (WCET) estimations are used. Dual criticality systems use a less pessimistic, but with lower level of assurance, WCET estimation, and a safe, but pessimistic, WCET estimation. Initially, both high and low criticality tasks are executed. When a high criticality task exceeds its less pessimistic WCET, the system switches mode and low criticality tasks are usually dropped, reducing the overall system Quality of Service (QoS). To postpone mode switch, and thus, improve QoS, existing approaches explore the slack, created dynamically, when the actual execution of a task is faster than its WCET. However, existing approaches observe this slack only after the task has finished execution. To enhance dynamic slack exploitation, we propose a fine-grained approach [19] that is able to expose the slack during the progress of a task, and safely uses it to postpone mode switch. The evaluation results show that the proposed approach has lower cost and achieves significant improvements in avoiding mode-switch, compared to existing approaches.

This work is done in collaboration with Stafanos Skalistis, Collins Aeropsace (Ireland).

7.7 Polyhedra at Work: Automatic Generation of VHDL Code for the Sherman-Morrison Formula

Participants: Patrice Quinton.

In collaboration with Université du Québec à Trois-Rivières (Canada) and Colorado State University (USA), the derivation of architectures for electrical simulation was studied. The problem is to compute in real-time the inverse of the admittance matrix modeling the circuit to simulate, in order to take into account possible changes of status of some switches in the circuit. This can be done using the Sherman-Morrison order one perturbation method. Architectures for this problem were generated using the MMAlpha synthesis tool. MMAlpha is a prototype software that was developped at Irisa for years to generate parallel architectures from a the Alpha, functional language. Another tool using this language, called AlphaZ, was also developped at CSU. With MMAlpha, Register Transfer Level (RTL) synthesizable code can be automatically produced and implemented on a FPGA platform. This research is described in [50].

7.8 Training Deep Neural Networks with Low-Precision Accelerators

Participants: Sami Ben Ali, 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.

Recent work has aimed to mitigate this computational challenge by introducing 8-bit floating-point (FP8) formats for multiplication. However, accumulations are still done in either half (16-bit) or single (32-bit) precision arithmetic. In [25], we investigate lowering accumulator word length while maintaining the same model accuracy. We present a multiply-accumulate (MAC) unit with FP8 multiplier inputs and FP12 accumulations, which leverages an optimized stochastic rounding (SR) implementation to mitigate swamping errors that commonly arise during low precision accumulations. We investigate the hardware implications and accuracy impact associated with varying the number of random bits used for rounding operations. We additionally attempt to reduce MAC area and power by proposing a new scheme to support SR in floating-point MAC and by removing support for subnormal values. Our optimized eager SR unit significantly reduces delay and area when compared to a classic lazy SR design. Moreover, when compared to MACs utilizing single-or half-precision adders, our design showcases notable savings in all metrics. Furthermore, our approach consistently maintains near baseline accuracy across a diverse range of computer vision tasks, making it a promising alternative for low-precision DNN training.

This work is conducted in collaboration with University of British Columbia, Vancouver, Canada.

7.9 Compression for DNN Inference

Participants: Thibault Allenet, Cédric Gernigon, Baptiste Rossigneux, Silviu Filip, Olivier Sentieys, Emmanuel Casseau.

Artificial intelligence (AI) on the edge has emerged as an important research area in the last decade to deploy different applications in the domains of computer vision and natural language processing on tiny devices. These devices have limited on-chip memory and are battery-powered. On the other hand, deep neural network (DNN) models require large memory to store model parameters and intermediate activation values. Thus, it is critical to make the models smaller so that their on-chip memory requirements are reduced.

In [17], we explore lossless DNN compression through exponent sharing. Various existing techniques like quantization and weight-sharing reduce model sizes at the expense of some loss in accuracy. We propose a lossless technique of model size reduction by focusing on the sharing of exponents in weights, which is different from the sharing of weights. We present results based on generalized matrix multiplication (GEMM) in DNN models. Our method achieves at least a 20% reduction in memory when using Bfloat16, and around 10% reduction when using IEEE single-precision floating point, with a very small impact (up to 10% on the processor and less than 1% on FPGA) on the execution time with no loss in accuracy. On specific models from HLS4ML, about 20% reduction in memory is observed in single precision with little execution overhead.

Quantization-Aware Training (QAT) has recently showed a lot of potential for low-bit settings in the context of image classification. Approaches based on QAT are using the Cross Entropy Loss function which is the reference loss function in this domain. In [53], we investigate quantization-aware training with disentangled loss functions. We qualify a loss to disentangle as it encourages the network output space to be easily discriminated with linear functions. We introduce a new method, Disentangled Loss Quantization Aware Training (DL-QAT), as our tool to empirically demonstrate that the quantization procedure benefits from those loss functions. Results show that the proposed method substantially reduces the loss in top-1 accuracy for low-bit quantization on CIFAR10, CIFAR100 and ImageNet. Our best result brings the top-1 Accuracy of a Resnet-18 from 63% to 64% with binary weights and 2-bit activations when trained on ImageNet. This work is conducted in collaboration with CEA List, Saclay.

One of the major bottlenecks in high-resolution Earth Observation (EO) space systems is the downlink between the satellite and the ground. Due to hardware limitations, onboard power limitations or ground-station operation costs, there is a strong need to reduce the amount of data transmitted. Various processing methods can be used to compress the data. One of them is the use of on-board deep learning to extract relevant information in the data. However, most ground-based deep neural network parameters

and computations are performed using single-precision floating-point arithmetic, which is not adapted to the context of on-board processing. In [30], we propose to rely on quantized neural networks and study how to combine low precision (mini) floating-point arithmetic with a Quantization-Aware Training methodology. We evaluate our approach with a semantic segmentation task for ship detection using satellite images from the Airbus Ship dataset. Our results show that 6-bit floating-point quantization for both weights and activations can compete with single-precision without significant accuracy degradation. Using a Thin U-Net 32 model, only a 0.3% accuracy degradation is observed with 6-bit minifloat quantization (a 6-bit equivalent integer-based approach leads to a 0.5% degradation). An initial hardware study also confirms the potential impact of such lowprecision floating-point designs, but further investigation at the scale of a full inference accelerator is needed before concluding whether they are relevant in a practical on-board scenario.

Deep learning edge devices face difficulties because of limiting factors, namely memory accesses and FLOPS due to the large number of parameters or the storing of activations. A current endeavor is to take a pretrained neural network and retrain a reduced version with as little computation overhead as possible. We are investigating a new approach for activation compression [44]. Our method involves interposing projection layers into a pretrained network around the nonlinearity, reducing the channel dimensionality through compression operations and then expanding it back. Our module is made to be then totally fused with the convolutions around it, guaranteeing no overhead, and maximum FLOPs reduction with low accuracy loss. Quantization impact on the compressed models is investigated. This work is done in collaboration with Inna Kucher and Vincent Lorrain, CEA List, Saclay.

7.10 Word-Length Optimization

Participants: Van-Phu Ha, 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.

With the growing complexity of applications, designers need to fit more and more computing kernels into a limited energy or area budget. Therefore, improving the quality of results of applications in electronic devices with a constraint on its cost is becoming a critical problem. Word Length Optimization (WLO) is the process of determining bit-width for variables or operations represented using fixed-point arithmetic to trade-off between quality and cost. State-of-the-art approaches mainly solve WLO given a quality (accuracy) constraint. In [33], we first show that existing WLO procedures are not adapted to solve the problem of optimizing accuracy given a cost constraint. It is then interesting and challenging to propose new methods to solve this problem. Then, we propose a Bayesian optimization based algorithm to maximize the quality of computations under a cost constraint (i.e., energy in this work). Experimental results indicate that our approach outperforms conventional WLO approaches by improving the quality of the solutions by more than 170%.

7.11 Machine-Efficient Rational Approximations of Mathematical Functions

Participants: Silviu Filip.

Software implementations of mathematical functions often use approximations that can be either polynomial or rational in nature. While polynomials are the preferred approximation in most cases, rational approximations are nevertheless an interesting alternative when dealing with functions that have a pronounced "nonpolynomial behavior" (such as poles close to the approximation domain, asymptotes or finite limits at $\pm\infty$). The major challenge is that of computing good rational approximations with machine number coefficients (e.g. floatingpoint or fixed-point) with respect to the supremum norm, a key step in most procedures for evaluating a mathematical function. This is made more complicated

by the fact that even when dealing with real-valued coefficients, optimal supremum norm solutions are sometimes difficult to obtain. In [27] , we introduce flexible and fast algorithms for computing such rational approximations with both real and machine number coefficients. Their effectiveness is explored on several examples.

This is joint work with Nicolas Brisebarre, École Normale Superieure de Lyon.

7.12 Exploiting Assertions Mining and Fault Analysis to Guide RTL-Level Approximation

Participants: Marcello Traiola.

In Approximate Computing (AxC), several design exploration approaches and metrics have been proposed to identify the approximation targets at the gate level, but only a few of them works on RTL descriptions. In addition, the possibility of combining the information derived from assertions and fault analysis is still under-explored. To fill in the gap, we propose an automatic methodology to guide the AxC design exploration at RTL [26]. Two approximation techniques are considered, bit-width reduction and statement reduction, and fault injection is used to mimic their effect on the design under exploration. Assertions are then dynamically mined from the original RTL description and the variation of their truth values is evaluated with respect to the injection of faults. These variations are then used to rank different approximation alternatives, according to their estimated impact on the functionality of the target design. The experiments, conducted on two modules widely used for image elaboration, show that the proposed approach represents a promising solution toward the automatization of AxC design exploration at RTL.

7.13 Automatic Approximation of Computer Systems Through Multi-objective Optimization

Participants: Marcello Traiola.

Given the escalating demands of contemporary applications for unparalleled computational resources, conventional paradigms in computing system design fall short in ensuring substantial performance improvements without escalating costs. To address this challenge, Approximate Computing (AxC) emerges as a promising solution, offering the potential to enhance computational performance by loosening stringent requirements on non-critical functional aspects of the system. In [52], we address the automatic approximation of computer systems through multi-objective optimization. Firstly, we present our automatic design methodology, i.e., how we model the approximate design space to be automatically explored. The exploration is achieved through multi-objective optimization to find good trade-offs between the system efficiency and accuracy. Then, we show how the methodology is applied to the systematic and application-independent design of generic combinational logic circuits, based on non-trivial local rewriting of and-inverter graphs (AIGs). Finally, to push forward the approximation limits, we showcase the design of approximate hardware accelerators for image processing and for common machine-learning-based classification models.

7.14 Input-Aware Accuracy Characterization for Approximate Circuits

Participants: Marcello Traiola.

Approximate Computing (AxC) can be applied systematically at various abstraction levels to increase the efficiency of several applications such as image processing and machine learning. Despite its benefit, AxC is still agnostic concerning the specific workload (i.e., input data to be processed) of a given application.

For instance, in signal processing applications (such as a filter), some inputs are constants (filter coefficients). Meaning that a further level of approximation can be introduced by considering the specific input distribution. This approach has been referred to as “input-aware approximation”. In [43], we explore how the input-aware approximate design approach can become part of a systematic, generic, and automatic design flow by knowing the data distribution. In particular, we show how input distribution can affect the error characteristics of an approximate arithmetic circuit and also the advantage of considering the data distribution by designing an input-aware approximate multiplier specifically intended for a high-pass FIR filter, where the coefficients are constant. Experimental results show that we can significantly reduce power consumption while keeping an error rate lower than state-of-the-art approximate multipliers.

7.15 Side-Channel Attacks on Embedded Artificial Intelligence

Participants: Benoit Coqueret, Olivier Sentieys.

Artificial intelligence, and specifically deep neural networks (DNNs), has rapidly emerged in the past decade as the standard for several tasks from specific advertising to object detection. The performance offered has led DNN algorithms to become a part of critical embed- ded systems, requiring both efficiency and reliability. In particular, DNNs are subject to malicious examples designed in a way to fool the network while being undetectable to the human observer: the adversarial examples. While previous studies propose frameworks to implement such attacks in black box settings, those often rely on the hypothesis that the attacker has access to the logits of the neural network, breaking the assumption of the traditional black box. In [28], we investigate a real black box scenario where the attacker has no access to the logits. In particular, we propose an architecture-agnostic attack which solve this constraint by extracting the logits. Our method combines hardware and software attacks, by performing a side-channel attack that exploits electromagnetic leakages to extract the logits for a given input, allowing an attacker to estimate the gradients and produce state-of-the-art adversarial examples to fool the targeted neural network. Through this example of adversarial attack, we demonstrate the effectiveness of logits extraction using side-channel as a first step for more general attack frameworks requiring either the logits or the confidence scores.

7.16 Algorithmic-Based Fault Detectors for Stencil Computations

Participants: Louis Narmour, Steven Derrien.

This work addresses the problem of transient errors detection in scientific computing, such as those occurring due to cosmic radiation or hardware component aging and degradation, using Algorithm-Based Fault Detection (ABFD). ABFD methods typically work by adding some additional computation in the form of invariant checksums which, by definition, should not change as the program executes. By computing and monitoring checksums, it is possible to detect errors by observing differences in the checksum values. However, this is challenging for two key reasons: (1) it requires careful manual analysis of the input program to infer a valid checksum expression, and (2) care must be taken to subsequently carry out the checksum computations efficiently enough for it to be worth it. Prior work has shown how to apply ABFT schemes with low overhead for a variety of input programs. In [38], we focus on a subclass of programs called stencil applications, which are an important class of computations found widely in various scientific computing domains.We have proposed a new compilation scheme and analysis to automatically analyze and generate the checksum computations.

7.17 Reliability Analysis and Evaluation

Participants: Fernando Fernandes Dos Santos, Angeliki Kritikakou, Marcello Traiola, Olivier Sentieys.

With the technology reduction, hardware resources became highly vulnerable to faults occurring even under normal operation conditions, which was not the case with technology used a decade ago. As a result, the evaluation of the reliability of hardware platforms is of highest importance. Such an evaluation can be achieved by exposing the platform to radiation and by forcing faults inside the system, usually through simulation-based methods.

As radiation-based reliability analysis can provide realistic and accurate reliability evaluation, we have performed several reliability analysis considering RISC-V based, GPU-based and FPGA-based platforms. More precisely, we characterize machine learning applications' vulnerabilities on RISC-V processors, by evaluating the neutron-induced error rate of Convolutional Neural Network (CNN) basic operations running on a RISC-V processor, GAP8 [47]. Our results show that executing the algorithm in parallel increases performance, and memory errors are the major contributors to the device error rate. Regarding GPUs, Vision Transformers (ViTs) are the new trend to improve performance and accuracy of machine learning. Through neutron beam experiments we show that ViTs have a higher FIT rate than traditional models but similar error criticality in [48]. Last, we characterize the impact of High-Level Synthesis (HLS) on the reliability of Neural Networks on FPGAs exposed to neutron. Our results show that the larger the circuit generated by HLS, the larger the error rate [51]. However, the larger accelerators can produce more correct inferences before experiencing failures.

7.18 harDNNing: a Machine-Learning-Based Framework for Fault-Tolerance Assessment and Protection of Deep Neural Networks

Participants: Marcello Traiola, Angeliki Kritikakou, Olivier Sentieys.

Deep Neural Networks (DNNs) show promising performance in several application domains, such as robotics, aerospace, smart healthcare, and autonomous driving. Nevertheless, DNN results may be incorrect, not only because of the network intrinsic inaccuracy, but also due to faults affecting the hardware. Indeed, hardware faults may impact the DNN inference process and lead to prediction failures. Therefore, ensuring the fault tolerance of DNN is crucial. However, common fault tolerance approaches are not cost-effective for DNNs protection, because of the prohibitive overheads due to the large size of DNNs and of the required memory for parameter storage. In [45, 46], we propose a comprehensive framework to assess the fault tolerance of DNNs and cost-effectively protect them. As a first step, the proposed framework performs datatype-andlayer-based fault injection, driven by the DNN characteristics. As a second step, it uses classification-based machine learning methods in order to predict the criticality, not only of network parameters, but also of their bits. Last, dedicated Error Correction Codes (ECCs) are selectively inserted to protect the critical parameters and bits, hence protecting the DNNs with low cost. Thanks to the proposed framework, we explored and protected eight representative Convolutional Neural Networks (CNNs). The results show that it is possible to protect the critical network parameters with selective ECCs while saving up to 83% memory w.r.t. conventional ECC approaches.

7.19 A Survey on Deep Learning Resilience Assessment Methodologies

Participants: Marcello Traiola.

Deep Learning (DL) applications are gaining increasing interest in the industry and academia for their outstanding computational capabilities. Indeed, they have found successful applications in various areas and domains such as avionics, robotics, automotive, medical wearable devices, gaming; some

have been labeled as safety-critical, as system failures can compromise human life. Consequently, DL reliability is becoming a growing concern, and efficient reliability assessment approaches are required to meet safety constraints. The article in [24] presents a survey of the main DL reliability assessment methodologies, focusing mainly on Fault Injection (FI) techniques used to evaluate the DL resilience. The article describes some of the most representative state-of-the-art academic and industrial works describing FI methodologies at different levels of abstraction. Finally, a discussion of the advantages and disadvantages of each methodology is proposed to provide valuable guidelines for carrying out safety analyses.

7.20 Fault-Tolerant Microarchitectures

Participants: Fernando Fernandes dos Santos, Romaric Nikiema, Angeliki Kritikakou, Marcello Traiola, Olivier Sentieys.

The migration of the computation from the cloud into edge devices, i.e., Internet-of-Things (IoTs) devices, reduces the latency and the quantity of data flowing into the network. With the emerging open-source and customizable RISC-V Instruction Set Architecture (ISA), cores based on such ISA are promising candidates for several application domains within the IoT family, such as automotive, Unnamed Aerial Vehicles (UAVs), industrial automation, healthcare, agriculture etc., where power consumption, real-time execution, security and reliability are of highest importance. In this emerging new era of connected RISC-V IoT devices, mechanisms are needed for a reliable and secure execution, still meeting area, energy consumption and computation time constraints of edge devices. We propose three mechanisms towards this goal [41], i.e., (i) a Root of Trust module for post-quantum secure boot, (ii) hardware checkers against hardware trojan horses and microarchitectural side-channel attacks, and (iii) a fine-grained dual core lockstep mechanism for real-time error detection and correction.

Real-time execution requires bounds in the worst-case execution time, while reliable execution is under threat, as systems are becoming more and more sensitive to transient faults. Thus, systems should be enhanced with fault-tolerant mechanisms with bounded error detection and correction overhead. Such mechanisms are typically based on redundancy at different granularity levels. Coarse-grained granularity has low comparison overhead, but may jeopardize timing guarantees. Fine-grained granularity immediately detects and corrects the error, but its implementation has increased design complexity. To mitigate this design complexity, we leverage high-level specification languages to design intrusive fine-grained lockstep processors based on the use of shadow registers and rollback, with bounded error detection and correction time, being appropriate for critical systems [40].

Furthermore, the majority of approaches, dealing with hardware faults, address the impact of faults on the functional behavior of an application, i.e., denial of service and binary correctness. Few approaches address the impact of faults on the application timing behavior, i.e., time to finish the application, and target faults occurring in memories. However, as the transistor size in modern technologies is significantly reduced, faults in cores cannot be considered negligible anymore. This work shows that faults not only affect the functional behavior, but they can have a significant impact on the timing behavior of applications. To expose the overall impact of faults, we enhance vulnerability analysis to include not only functional, but also timing correctness, and show that faults impact WCET estimations [39]. A RISC-V core is used as a case study. The obtained results show that faults can lead up to almost 700% increase in the maximum observed execution time between fault-free and faulty execution without protection, affecting the WCET estimations.

7.21 Fault-Tolerant Networks-on-Chip

Participants: Wilfread Guilleme, Hamza Amara, Ibrahim Krayem, Cédric Killian, Angeliki Kritikakou, Emmanuel Casseau, 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. At the same time, artificial intelligent algorithms are deployed in large domains of applications, and particularly in embedded systems to support edge computing and limit data transfer toward the cloud. For embedded Neural Networks (NNs), faults, such as Single-Event Upsets (SEUs), may have a great impact on their reliability, [32]. To address this challenge, previous works have been done about SEU layers sensitivity of AI models. Contrary to these techniques, remaining at high level, we propose a more accurate analysis, highlighting that faults transitioning from 0 to 1 significantly impact classification outcomes. Based on this specific behavior, we propose a simple hardware block able to detect and mitigate the SEU impact. Obtained results show that HTAG (Hardening Technique with And Gates) protection efficiency is near 96% for the LeNet-5 CNN inference model, suitable for an embedded system. This result can be improved with other protection methods for the classification layer. Additionally, it significantly reduces area overhead and critical path compared to existing approaches.

Furthermore, as faults can be the consequence of intentional attack on the NoC, we also study their impact on data transfer when a specific compression format is applied on NoC traffic. More specifically, taking the transfer of an image as a use case, we study the degradations induced by an attack on the payload of the data traffic for non compressed data and also for FlitZip compressed data [59]. Experiments shows that mean square error of uncompressed data increases more rapidly with fault injection rate than for FlitZip-based approach. In fact, mean square error difference between uncompressed and compressed approaches increases with fault injection rate increasing. Because header flit of each packet is very sensitive to faults, we also propose an extension of FlitZip technique that makes it possible to include protection into the header flit without increasing the number of bits of the header.

7.22 Fault-Tolerant Task Deployment onto Multicore Systems

Participants: Emmanuel Casseau, 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 reduce energy consumption, while enhancing reliability and satisfying real-time constraints, we propose a hybrid approach that combines distinct reliability enhancement techniques, under task-level, processor-level and system-level DVFS. Our task mapping problem jointly decides task allocation, task frequency assignment, and task duplication, under real-time and reliability constraints. This is achieved by formulating the task mapping problem as a Mixed Integer Non-Linear Programming problem, and equivalently transforming it into a Mixed Integer Linear Programming, that can be optimally solved [12].

Energy efficiency, real-time response, and data transmission reliability are important objectives during networked systems design. In [22], we developped an efficient task mapping scheme to balance these important but conflicting objectives. To achieve this goal, tasks are triplicated to enhance reliability and mapped on the wireless nodes of the networked systems with Dynamic Voltage and Frequency Scaling (DVFS) capabilities to reduce energy consumption while still meeting real-time constraints. Our contributions include the mathematical formulation of this task mapping problem as mixed-integer programming that balances node energy consumption, enhancing data reliability, under real-time and energy constraints. Compared with the State-of-the-Art (SoA), a joint-design problem is considered, where DVFS, task triplication, task allocation, and task scheduling are optimized concurrently. To find the optimal solution, the original problem is linearized, and a decomposition-based method is proposed. The optimality of the proposed method is proved rigorously. Furthermore, a heuristic based on the greedy algorithm is designed to reduce the computation time. The proposed methods are evaluated and compared through a series of simulations. The results show that the proposed triplication-based task mapping method on average achieves 24.84% runtime reduction and 28.62% energy saving compared to the SoA method

This work is done in collaboration with Lei Mo School of Automation, Southeast University (China).

7.23 Analytical Model for NoC Performance Analysis

Participants: Ibrahim Krayem, Cédric Killian, Daniel Chillet.

As the complexity of system-on-chip continues to increase, along with a growing number of heterogeneous cores and accelerators, evaluating architectural performance becomes a paramount concern to efficiently explore the design space. The on-chip interconnect in these architectures plays a crucial role, serving as the backbone for communication and thus influencing the overall system performance. In recent years, we have observed the emergence of 3D architectures based on chiplets. These advancements also enable the integration of multiple interconnects with various technologies, further intensifying the design complexity and evaluation. To circumvent this problem, we have developed an analytical method able to evaluate the performance of such heterogeneous interconnects. The method is based on queuing theory and computes latency of communication with respect to injected traffics [18, 35]. The proposed method is open source and available on gitlab.

This method is based on traffic models (in particular on packet injection rate defined by mathematical law) which not always models real application traffics. Consequently, we extend our method with a pre-computation step to analyze the NoC traffic trace from real applications. The goal of this step consists in identifying how real trace can be cut in several windows following a Poisson law. This step is managed by the accuracy targeted by the designer. The results demonstrate that our proposed model significantly reduces the simulation execution time by up to 500× while maintaining an error rate of less than 5% compared to the Noxim cycle-accurate simulator.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Participants: Olivier Sentieys, Joseph Paturel, Emmanuel Casseau, François Charot, Cédric Killian, Daniel Chillet.

Collaboration 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.

8.2 Bilateral Grants with Industry

Participants: Olivier Sentieys, Léo Pradels, Daniel Chillet, Silviu-Ioan Filip.

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.

Participants: Olivier Sentieys, Benoit Coqueret.

Thales is funding a PhD on physical security attacks against Artificial Intelligence based algorithms.

Participants: Daniel Chillet.

Orange Labs is funding a PhD on energy estimation of applications running on cloud. The goal is to analyze application profiles and to develop an accurate estimator of power consumption based on a selected subset of processor events.

Participants: Cédric Gernigon, Seungah Lee, Olivier Sentieys, Silviu-Ioan Filip, Angeliki Kritikakou, Emmanuel Casseau.

CNES is co-funding the PhD thesis of Cédric Gernigon on highly compressed/quantized neural networks for FPGA on-board processing in Earth observation by satellite, and the PhD thesis of Seungah Lee on efficient designs of on-board heterogeneous embedded systems for space applications.

Participants: Bertrand Le Gal, Simon Rokicki.

KeySom is funding the PhD thesis of Léo Pajot on efficient implementation of parallel applications such as CNN on custom RISC-V processor cores. The goal is to propose a CGRA like architecture and its compilation framework to ease platform designer work in accelerating developed systems.

Participants: Daniel Chillet.

Orange Labs is funding a PhD on energy estimation of applications running on cloud. The goal is to analyze application profiles and to develop an accurate estimator of power consumption based on a selected subset of processor events.

8.3 Informal Collaborations with Industry

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

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria Associate Team

EdgeTrain

Participants: Silviu-Ioan Filip, Olivier Sentieys, Guy Lemieux (UBC).

Title: Low-Precision Accelerators for Deep Learning Training on Edge Devices

Duration: 2022 - 2024

Coordinator: Silviu-Ioan Filip

Partners: University of British Columbia, Vancouver (Canada)

Summary: The main scientific objectives of the proposed collaborative research project are: (i) the analysis and development of custom arithmetic operators for DNN training acceleration and a working prototype accelerator for edge training; (ii) a design space exploration of the accelerators with respect to energy and power consumption by examining the number system(s) and bit widths used; the production of an automated design flow for the generation of custom accelerators targeting Field Programmable Gate Array (FPGA) Systems on Chip (SoC), specialized for a given deep neural network model to train.

9.1.2 Participation in other International Programs

IntelliVIS

Participants: Olivier Sentieys, Sharad Sinha (IIT Goa).

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

Duration: 2019 - 2023

Partner Institution: IIT Goa (India)

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.

LRS

Participants: Steven Derrien, Louis Narmour, Corentin Ferry, Sanjay Rajopadhye (CSU).

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).

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.
- University of Trento (Italy), Reliability analysis and radiation experiments
- 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.
- Department of Electrical and Computer Engineering, University of Naples (Italy), Digital Hardware Design Space Exploration for Approximate-Computing-based Applications
- Department of Control and Computer Engineering, Politecnico di Torino (Italy), Fault tolerance of Deep Neural Network hardware accelerators
- Department of Computer Science, University of Verona (Italy), Assertion-driven Design Exploration of Approximate Hardware

9.2 International research visitors

9.2.1 Visits of international scientists

Hendrik Wohrle, Professor at FH Dortmund, has visited TARAN for three months from Apr. 2023 until June 2023.

Louis Narmour, PhD Student from Colorado State University (CSU), USA, is visiting 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.

Jinyi Xu, PhD Student from East China Normal University, China, has visited TARAN from Nov. 2020 for two years.

Pavitra Bhade, PhD Student from IIT Goa, has visited TARAN for one month in Feb. 2023.

9.2.2 Visits to international teams

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

9.3 National initiatives

9.3.1 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.

9.3.2 ANR RAKES

Participants: Olivier Sentieys, Cédric Killian, Abhijit Das.

- 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 2024
- 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.

9.3.3 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 - May 2024
- 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.

9.3.4 ANR SHNOC

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

- 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.

9.3.5 ANR FASY

Participants: Angeliki Kritikakou, Marcello Traiola, 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.

9.3.6 ANR Re-Trusting

Participants: Olivier Sentieys, Angeliki Kritikakou, Marcello Traiola, 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.

9.3.7 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.

9.3.8 ANR LOTR

Participants: Steven Derrien, Simon Rokicki.

- Program: ANR PRC
- Project acronym: LOTR
- Project title: Lord Of The RISCs
- Duration: Oct. 2023 - Sep. 2027
- Coordinator: Steven Derrien
- Other partners: CEA, TARAN, PACAP

Lord Of The RISCs (LOTR) is a novel flow for designing highly customized RISC-V processor microarchitectures for embedded and IoT platforms. The LOTR flow operates on a description of the processor Instruction Set Architecture (ISA). It can automatically infer synthesizable Register Transfer Level (RTL) descriptions of a large number of microarchitecture variants with different performance/cost trade-offs. In addition, the flow integrates two domain-specific toolboxes dedicated to the support of timing predictability (for safety-critical systems) and security (through hardware protection mechanisms).

9.3.9 CYBERPROS

Participants: Olivier Sentieys.

- Program: BPI France
- Project title: Fault Injection Emulator for Cyberattacks and System Security Evaluation processeurs
- Duration: Oct. 2023 - Sep. 2026
- Coordinator: Patrice Deroux-Dauphin
- Other partners: TEMENTO, IROC

The objective of the CYBERPROS project is to be able to predict the behavior of a circuit subjected to to cyberattacks by fault injection. The research work consists of developing a active attack emulator and associated simulation tools. A hardened processor core will be developed as a test vehicle. Test results will be digitized for editing of learning algorithms underlying the creation of a database and tools for predictive behavior.

9.3.10 PEPR ARSENE

Participants: Louis Savary, Herinomena Andrianatrehina, Simon Rokicki, Steven Derrien, Ronan Lashermes, Olivier Sentieys.

- Program: PEPR Cyber
- Project title: Secure architectures for embedded digital systems
- Duration: Jul. 2022 - Jun. 2028
- Coordinator: CEA
- Other partners: CEA, PACAP, TARAN, LHC, Lab-STICC, LIRMM, Verimag, TIMA, LCIS, EMSE, Telecom Paris

The main objectives of the ARSENE project are to allow the French community to make significant advances in the field to strengthen the community's expertise and visibility on the international stage. Taran's contribution is on the study and implementation of two families of RISC-V processors: 32-bit RISC-V for low power secure circuits against physical attacks for IoT applications and 64-bit RISC-V secure circuits against micro-architectural attacks.

9.3.11 ANR RADYAL

Participants: Marcello Traiola, Olivier Sentieys.

- Program: ANR PRC
- Project acronym: RADYAL
- Project title: Resource-Aware DYnamically Adaptable machine Learning
- Duration: Oct 2023 – Apr 2027
- Coordinator: Stefan Duffner, LIRIS, Lyon
- Other partners: TARAN, LIRIS, CTRL-A (Inria Grenoble), GIPSA-LAB

Nowadays, for many applications, the performance requirements of a DNN model deployed on a given hardware platform are not static but evolving dynamically as its operating conditions and environment change. RADYAL studies original interdisciplinary approaches that allow DNN models to be dynamically configurable at run-time on a given reconfigurable hardware accelerator architecture, depending on the external environment, following an approach based on feedback loops and control theory.

9.3.12 ANR SEC-V

Participants: Bertrand LE GAL.

- Program: ANR PRCE
- Project acronym: SEC-V
- Project title: open-source, secure and high-performance processor core based on the RISC-V ISA
- Duration: Oct 2021 – Apr 2025

- Coordinator: Sebastien Pillement, IETR, Nantes
- Other partners: TARAN, LS2N, THALES TRT, THALES INVIA

In recent years, attacks exploiting optimization mechanisms have appeared. Exploiting, for example, flaws in cache memories, performance counters or speculation units, they call into question the safety and security of processors and the industrial systems that use them. SEC-V studies original interdisciplinary approaches that rely on RISC-V open-hardware architectures and CISC paradigm to prodive runtime flexibility and adaptability. The originality of the approach lies in the integration of a dynamic code transformation unit covering 4 of the 5 NIST functions of cybersecurity, notably via monitoring (identify, detect), obfuscation (protect), and dynamic adaptation (react). This dynamic management paves the way for on-line optimizations to improve the security and safety of the microarchitecture, without reworking either the software or the chip architecture.

10 Dissemination

Participants: Emmanuel Casseau, François Charot, Daniel Chillet, Steven Derrien, Fernando Fernandes Dos Santos, Silviu-Ioan Filip, Angeliki Kritikakou, Bertrand Le Gal, Simon Rokicki, Olivier Sentieys, Marcello Traiola.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- M. Traiola was the Vice General co-chair of IEEE IOLTS 2023.
- M. Traiola was the General co-chair of IEEE SELSE workshop 2023.
- A. Kritikakou was the Vice General co-chair of IEEE IOLTS 2023.

Member of the organizing committees

- D. Chillet was in the Organizing Committee of Rapido 2023.
- D. Chillet was the co-program Session-Chair of Gretsi 2023.
- O. Sentieys and M. Traiola were members of the IEEE/ACM DATE Executive Committee, 2023.
- M. Traiola was the Review Chair of IEEE/ACM DATE 2023.
- M. Traiola was the Publication Chair and Media Chair of IEEE VTS 2023.
- M. Traiola was the Publication Co-Chair of IEEE DFT 2023.
- A. Kritikakou was the Hot Topic Chair of IEEE RTSS 2023.
- A. Kritikakou was the PhD Topic Co-Chair of IEEE ETS 2023.
- A. Kritikakou was the Brief Presentation Co-Chair of IEEE RTAS 2023.

10.1.2 Scientific events: selection

Chair of conference program committees

- A. Kritikakou was the Program co-chair of IEEE SELSE workshop 2023.
- O. Sentieys was co-chair of the Focus Sessions at IEEE/ACM DATE Executive Committee, 2023.

Member of the conference program committees

- D. Chillet was member of the technical program committee of HiPEAC Rapido, HiPEAC WRC, DSD, ComPAS, DASIP, ARC.
- S. Filip was member of the technical program comittee for ARITH.
- M. Traiola was member of the technical program comittee for IEEE VTS, IEEE ETS, IEEE IOLTS, IEEE LATS, IEEE/ITRI VLSI-DAT, IEEE SELSE workshop, IEEE ARTS and eARTS workshops, Approximate Computing (AxC) workshop.
- F. Fernandes dos Santos was member of the technical program comittee for the IEEE IOLTS and IEEE SELSE.
- A. Kritikakou was member of the technical program committee of DATE, ETS, RTSS, ECRTS, RTAS, EMSOFT, RTCSA, ICPADS, ISVLSI, DS-RT, SAMOS, RTNS, ARC, COMPAS.
- O. Sentieys served as a committee member in the IEEE EDAA Outstanding Dissertations Award (ODA) 2023.
- O. Sentieys was a member of technical program committee of IEEE/ACM ICCAD, IEEE FPL, ACM ENSSys, ACM SBCCI, ARC.

10.1.3 Member of the editorial boards of Journals

- D. Chillet is member of the Editor Board of Journal of Real-Time Image Processing (JRTIP).
- A. Kritikakou is Handling Editor for Elsevier Microprocessors and Microsystems Journal.
- A. Kritikaou is Associate editor in Elsevier Journal of Systems Architecture.
- O. Sentieys is member of the editorial board of Journal of Low Power Electronics.

Reviewing activities

- M. Traiola was a reviewer for IEEE (ToC, TCAD, TECT, TCAS) and ACM journals (TECS, JETC, JATS, TODAES)
- F. Fernandes dos Santos was a reviewer for IEEE Transactions on Nuclear Science and Journal of Systems Architecture, Journal of Supercomputing, Microelectronics Reliability, and Microprocessors and Microsystems.
- D. Chillet was a reviewer for Microprocessors and Microsystems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Journal of Systems Architecture, and ACM Transactions on Architecture and Code Optimization.
- B. Le Gal was reviewer for IEEE Wireless Communications Letters, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Transactions on Circuits and Systems II: Express Briefs, IEEE Signal Processing Letters, IEEE Communications Letters, IEEE Signal Processing Letters, and IEEE Transactions on Image Processing.
- O. Sentieys was a reviewer for IEEE Transactions on Computers, and ACM Transactions on Architecture and Code Optimization.

10.1.4 Invited talks

- M. Traiola was an invited speaker at FETCH 2023.
- M. Traiola was an invited panelist at the 8th Workshop on Approximate Computing (AxC23) @ IEEE/IFIP DSN 2023.
- A. Kritikakou gave an invited talk on ”Reliability analysis and protection of RISC-V processors for safety-critical embedded systems” at BITFLIP workshop during the European Cyber Week (ECW) 2023.
- A. Kritikakou gave an invited talk on ”Reliability analysis and protection for AI critical embedded systems” at BarCamp x-GDR 2023 on implementation challenges of AI.
- O. Sentieys gave an invited talk at the European Nanoelectronics Applications Design and Technology Conference (ADTC) on ”Design and Exploration of RISC-V Cores from High-Level Specifications”.
- O. Sentieys gave an invited presentation at the Inria Scientific Days on ”Computing beyond Moore's Law”.
- S. Derrien gave an invited presentation at the Inria Scientific Days on ”How to (not) blow-up a pipeline”.

10.1.5 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 SoC2.
- A. Kritikakou is a member of the French National University Council in Computer Science (CNU - Conseil National des Universites, 27ème section) since 2022.
- A. Kritikakou is co-animator of the "High performance embedded computing" topic of GDR SoC2.
- 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 SoC2.
- O. Sentieys is an elected member of the Evaluation Committee (CE) of Inria.

10.1.6 Scientific expertise

- D. Chillet was a member of the HCERES Evaluation Committee for Master Programs of Université Nice Côte d'Azur.
- O. Sentieys was a member of the ANR Scientific Evaluation Committee CE25 "Software science and engineering - Multi-purpose communication networks, high-performance infrastructure".

10.1.7 Research administration

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

10.1.8 Standardization activities

- S. Filip and O. Sentieys are members of the IEEE P3109 Standardization Group on Arithmetic Formats for Machine Learning.

10.2 Teaching - Supervision - Juries

- A. Kritikakou is a member of the Examination Committee of Industrial Engineering Sciences and Computer Engineering (SII) Aggregation.

10.2.1 Teaching administration

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

10.2.2 Teaching

- E. Casseau: programmable logic, 30h, ENSSAT (L3)
- E. Casseau: low power design, 8h, ENSSAT (M1)
- E. Casseau: real time design methodology, 54h, ENSSAT (M1)
- E. Casseau: computer architecture, 24h, ENSSAT (M1)
- E. Casseau: VHDL design, 42h, ENSSAT (M1)
- E. Casseau: SoC and high-level synthesis, 24h, ENSSAT (M2)
- D. Chillet: embedded processor architecture, 20h, Enssat (M1)
- D. Chillet: multimedia processor architectures, 30h, Enssat (M2)
- D. Chillet: advanced processor architectures, 20h, Enssat (M2)
- D. Chillet: advanced processor architectures, 15h, Embedded Systems Master (M2)
- D. Chillet: micro-controller, 32h, Enssat (L3)
- D. Chillet: low-power digital CMOS circuits, 4h, UBO (M2)
- D. Chillet: Processor architectures, 40h, USTH (B2)
- 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. Filip: Operating Systems, 24h, ENS Rennes (Aggregation Informatique)
- A. Kritikakou: Tools and programming in C, 24.75h, istic (L3)
- A. Kritikakou: Computer programming, 22.5h, istic (L3)
- A. Kritikakou: Unix commands and programming, 6.75h, istic (L3)
- A. Kritikakou: Fault tolerant embedded systems, 6h, INSA (M2)

- A. Kritikakou: Energy sobriety of digital architectures, 7.5h, INSA (M2)
- S. Rokicki: C Programming, 24h, ENS Rennes
- O. Sentieys: VLSI integrated circuit design, 24h, ENSSAT (M1)
- O. Sentieys: VHDL and logic synthesis, 18h, ENSSAT (M1)
- O. Sentieys: Hardware Accelerators for Deep Neural Networks, 54h, Master of Embedded Systems, ISTIC (M2)
- M. Traiola: Operating Systems, 24h, ENS Rennes (Aggregation Mecatronique)

10.2.3 PhD Supervision

- PhD: Thibault Allenet, Quantization and adversarial robustness of embedded deep neural networks, March 2023, O. Sentieys, O. Bichler (CEA LIST) [53].
- PhD: Minh Thanh Cong, Hardware accelerated simulation and automatic design of heterogeneous architecture, March 2023, F. Charot, S. Derrien [54].
- PhD: Van-Phu Ha, Contributions to the scalability of automatic precision tuning, March 2023, O. Sentieys [55].
- PhD in progress: Hamza Amara, Detection and countermeasures for DoS attack in Noc-based SoC using machine learning, Oct. 2022, E. Casseau, D. Chillet, C. Killian.
- PhD in progress: Herinomena Andrianatrehina, Ensuring confidentiality in modern Out-of-Order cores, Nov 2022, S. Rokicki, R. Lashermes.
- PhD in progress: Gaetan Barret, Predictive model of energy consumption of cloud-native applications, Nov. 2022, D. Chillet.
- PhD in progress: Sami Ben Ali, Efficient Low-Precision Training for Deep Learning Accelerators, Jan. 2022, O. Sentieys, S. Filip.
- PhD in progress: Benoit Coqueret, Physical Security Attacks Against Artificial Intelligence Based Algorithms, CIFRE Thesis with Thales, Nov. 2022, O. Sentieys, M. Carbone (Thales), G. Zaid (Thales).
- PhD in progress: Léo De La Fuente, In-Memory Computing for Ultra Low Power Architectures, Nov. 2021, O. Sentieys, J.-F. Christmann (CEA).
- PhD in progress: Paul Estano, Dynamic Precision Training of Deep Neural Network Models on the Edge, Feb. 2022, S. Filip, E. Riccietti (ENS Lyon), S. Derrien.
- 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: 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: Wilfread Guilleme, Fault Tolerant Hardware Architectures for Artificial Intelligence, Oct. 2022, D. Chillet, C. Killian, A.Kritikakou.
- PhD in progress: Ibrahim Krayem, Fault tolerant emerging on-chip interconnects for manycore architectures, Oct. 2020, C. Killian, D. Chillet.
- PhD in progress: Seungah Lee, Efficient Designs of On-Board Heterogeneous Embedded Systems for Space Applications, Nov. 2021, A. Kritikakou, E. Casseau, R. Salvador, O. Sentieys.
- PhD in progress: Dylan Leothaud, Automatic synthesis of secure and predictable processors for the Internet of Thing, Oct. 2023, S. Derrien, S. Rokicki.
- PhD in progress: Guillaume Lomet, Guess What I'm Learning: Side-Channel Analysis of Edge AI Training Accelerators, Oct. 2022, C. Killian, R. Salvador, 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: 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 progrees: Romaric (Pegdwende) Nikiema, Time-guaranteed and reliable execution for real-time multicore architectures, Oct. 2022, A. Kritikakou, M. Traiola
- PhD in progress: Leo Pajot, Soft-core processor with dynamic binary execution exploiting instruction-level parallelism, CIFRE Thesis with KEYSOM SAS, Sep. 2023, B. Le Gal.
- PhD in progress: Leo Pradels, Constrained optimization of FPGA accelerators for embedded deep convolutional neural networks, CIFRE Thesis with Safran, Dec. 2020, D. Chillet, O. Sentieys, S. Filip.
- PhD in progress: Baptiste Rossigneux, Adapting sparsity to hardware in neural networks, Nov. 2022, E. Casseau, I. Kucher (CEA), V. Lorrain (CEA).
- PhD in progress: Louis Savary, Security of DBT-based processors, Sept 2022, S. Rokicki, S. Derrien.

10.3 Popularization

10.3.1 Interventions

As part of a collaboration between multiple institutes and universities (Inria Rennes, the Federal University of Rio Grande do Sul, and the University of Trento), we helped develop a demonstration showcasing radiation-induced faults' impacts on Deep Neural Networks (DNNs) at the ChipIr Facility in the Rutherford Appleton Laboratory, located in Didcot, UK. The demonstration consists of an NVIDIA Jetson Orin board running a DNN and performing semantic segmentation on a self-driving car video of London. A large red button is placed to simulate a radiation-induced fault occurring during the DNN execution. Users can see in real-time how an unprotected DNN could affect the reliability of an autonomous car by just pushing the red button. This platform is used for demonstrations to high school and college students from the Oxford region while visiting the facility.

11 Scientific production

11.1 Major publications

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] A. Ruospo, E. Sanchez, L. Matana Luza, L. Dilillo, M. Traiola and A. Bosio. 'A Survey on Deep Learning Resilience Assessment Methodologies'. In: *Computer* 56 (Feb. 2023), pp. 57–66. DOI: 10.1109/MC.2022.3217841. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03834128.

11.2 Publications of the year

International journals

[12] M. Cui, A. Kritikakou, L. Mo and E. Casseau. 'Near-Optimal Energy-Efficient Partial-Duplication Task Mapping of Real-Time Parallel Applications'. In: *Journal of Systems Architecture* 134 (Jan. 2023), p. 102790. DOI: 10.1016/j.sysarc.2022.102790. URL: https://hal.science/hal-03888480.

[13] P. Dobiáš, E. Casseau and O. Sinnen. 'Online Fault Tolerant Energy-Aware Algorithm for CubeSats'. In: *Sustainable Computing : Informatics and Systems* (9th Mar. 2023), p. 100853. DOI: 10.1016/j.suscom.2023.100853. URL: https://hal.science/hal-04033761.

[14] F. Fernandes dos Santos, L. Carro, F. Vella and P. Rech. 'Assessing the Impact of Compiler Optimizations on GPUs Reliability'. In: *ACM Transactions on Architecture and Code Optimization* (12th Jan. 2024). DOI: 10.1145/3638249. URL: https://hal.science/hal-04398273.

[15] C. Ferry, T. Yuki, S. Derrien and S. Rajopadhye. 'Increasing FPGA Accelerators Memory Bandwidth with a Burst-Friendly Memory Layout'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2023), pp. 1–15. DOI: 10.1109/TCAD.2022.3201494. URL: https://inria.hal.science/hal-03930715.

[16] C. Gillet, A. Vincent, B. Le Gal and S. Saïghi. 'A High-Level Methodology to Evaluate and Optimize Digital Architectures Targeting Spike Encoding'. In: *IEEE Access* 11 (2023), pp. 120654–120665. DOI: 10.1109/ACCESS.2023.3324877. URL: https://hal.science/hal-04406432.

[17] P. Kashikar, O. Sentieys and S. Sinha. 'Lossless Neural Network Model Compression Through Exponent Sharing'. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 31 (Nov. 2023), pp. 1816–1825. DOI: 10.1109/tvlsi.2023.3307607. URL: https://hal.science/hal-04397024.

[18] I. Krayem, J. O. Sosa, C. Killian and D. Chillet. 'Analytical Model for Performance Evaluation of Token-Passing-Based WiNoCs'. In: *IEEE Design & Test* 40.6 (Aug. 2023), pp. 136–148. DOI: 10.1109/MDAT.2023.3309730. URL: https://inria.hal.science/hal-04373575.

[19] A. Kritikakou and S. Skalistis. 'Mitigating Mode-Switch through Run-time Computation of Response Time'. In: *ACM Transactions on Design Automation of Electronic Systems* 28.5 (9th Sept. 2023), pp. 1–26. DOI: 10.1145/3597432. URL: https://hal.science/hal-04397350.

[20] B. Le Gal, C. Jego and V. Pignoly. 'High-performance hard-input LDPC decoding on multi-core devices for optical space links'. In: *Journal of Systems Architecture* 137 (Apr. 2023), p. 102832. DOI: 10.1016/j.sysarc.2023.102832. URL: https://hal.science/hal-04406492.

[21] L. Mo, X. Li, A. Kritikakou and P. You. 'Optimal IC Task Mapping to Maximize QoS on Heterogeneous Multicore Systems'. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* (1st Nov. 2023), pp. 1–5. DOI: 10.1109/TCSII.2023.3329050. URL: https://hal.science/hal-04267659.

[22] L. Mo, Q. Zhou, A. Kritikakou and X. Cao. 'Energy Optimized Task Mapping for Reliable and Real-Time Networked Systems'. In: *ACM Transactions on Sensor Networks* (21st Feb. 2023), pp. 1–24. DOI: 10.1145/3584985. URL: https://hal.science/hal-03999363.

[23] C. Monière, B. Le Gal and E. Boutillon. 'Real-time energy-efficient software and hardware implementations of a QCSP communication system'. In: *Journal of Systems Architecture* 141 (Aug. 2023), p. 102933. DOI: 10.1016/j.sysarc.2023.102933. URL: https://hal.science/hal-04406481.

[24] A. Ruospo, E. Sanchez, L. Matana Luza, L. Dilillo, M. Traiola and A. Bosio. 'A Survey on Deep Learning Resilience Assessment Methodologies'. In: *Computer* 56 (Feb. 2023), pp. 57–66. DOI: 10.1109/MC.2022.3217841. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03834128.

International peer-reviewed conferences

[25] S. Ben Ali, S.-I. Filip and O. Sentieys. 'A Stochastic Rounding-Enabled Low-Precision Floating-Point MAC for DNN Training'. In: 27th IEEE/ACM Design, Automation and Test in Europe (DATE). Valencia, Spain, 25th Mar. 2024, pp. 1–6. URL: https://hal.science/hal-04380270.

[26] A. Bosio, S. Germiniani, G. Pravadelli and M. Traiola. 'Exploiting assertions mining and fault analysis to guide RTL-level approximation'. In: DATE 2023 – 26th IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium, 17th Apr. 2023. URL: https://inria.hal.science/hal-03887685.

[27] ***Best Paper***
N. Brisebarre and S.-I. Filip. 'Towards Machine-Efficient Rational L ∞ -Approximations of Mathematical Functions'. In: *Proceedings of the 30th IEEE International Symposium on Computer Arithmetic ARITH 2023, Sep 2023, Portland, Oregon, USA*. 30th IEEE International Symposium on Computer Arithmetic ARITH 2023. Portland, United States, 4th Sept. 2023. URL: https://hal.science/hal-04093020.

[28] B. Coqueret, M. Carbone, O. Sentieys and G. Zaid. 'When Side-Channel Attacks Break the Black-Box Property of Embedded Artificial Intelligence'. In: AISec 2023 - 16th ACM Workshop on Artificial Intelligence and Security. Copenhagen, Denmark: ACM, 26th Nov. 2023, pp. 127–138. DOI: 10.1145/3605764.3623903. URL: https://hal.science/hal-04320434.

[29] F. Fernandes dos Santos, L. Carro and P. Rech. 'Understanding and Improving GPUs' Reliability Combining Beam Experiments with Fault Simulation'. In: ITC 2023 - IEEE International Test Conference. Anaheim, United States: IEEE, 2023, pp. 176–185. DOI: 10.1109/ITC51656.2023.00034. URL: https://hal.science/hal-04398063.

[30] C. Gernigon, S.-I. Filip, O. Sentieys, C. Coggiola and M. Bruno. 'Low-Precision Floating-Point for Efficient On-Board Deep Neural Network Processing'. In: European Data Handling & Data Processing Conference (EDHPC). Juan-Les-Pins, France, Oct. 2023, pp. 1–8. URL: https://hal.science/hal-04252197.

[31] J.-M. Gorius, S. Rokicki and S. Derrien. 'A Unified Memory Dependency Framework for Speculative High-Level Synthesis'. In: CC'24 - ACM SIGPLAN 2024 International Conference on Compiler Construction. Edinburgh (Ecosse), United Kingdom, 2nd Mar. 2024. URL: https://inria.hal.science/hal-04394762.

[32] W. Guilleme, Y. Helen, R. Priem, A. Kritikakou, D. Chillet and C. Killian. 'Protection sélective d'un réseau de neurones implémenté sur puce FPGA soumis à un environnement radiatif'. In: Gretsi 2023 - XXIXème Colloque Francophone de Traitement du Signal et des Images. Grenoble, France, 2023, pp. 1–4. URL: https://inria.hal.science/hal-04396267.

[33] V.-P. Ha and O. Sentieys. 'Maximizing Computing Accuracy on Resource-Constrained Architectures'. In: DATE 2023 - 26th IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium: IEEE, 17th Apr. 2023, pp. 1–6. URL: https://inria.hal.science/hal-03885240.

[34] S. S. Hoseininasab, C. Collange and S. Derrien. 'Rapid Prototyping of Complex Micro-architectures Through High-Level Synthesis'. In: *Applied Reconfigurable Computing*. ARC 2023 - 19th International Symposium on Applied Reconfigurable Computing. Vol. 14251. Lecture Notes in Computer Science. Cottbus, Germany: Springer Nature Switzerland, 2023, pp. 19–34. DOI: 10.1007/978-3-031-42921-7_2. URL: https://hal.science/hal-04225360.

[35] I. Krayem, J. O. Sosa, C. Killian and D. Chillet. 'Analytical Model for Performance Evaluation of Token-Passing Based WiNoCs'. In: NOCS 2023 - 17th IEEE/ACM International Symposium on Networks-on-Chip. Hamburg, Germany, 2023. URL: https://inria.hal.science/hal-04373591.

[36] S. Lee, R. Salvador, A. Kritikakou, O. Sentieys, J. Galizzi and E. Casseau. 'High-Level Synthesis-Based On-board Payload Data Processing considering the Roofline Model'. In: *EDHPC 2023 proceedings*. EDHPC 2023 - European Data Handling & Data Processing Conference. Juan-Les-Pins, France, 2nd Oct. 2023, pp. 1–10. URL: https://inria.hal.science/hal-04294305.

[37] M. Magnant, M. A. Ben Temim, B. Le Gal, G. Ferré and F. Collard. 'Implantation d'un détecteur de préambules vobulés par satellite en orbite basse'. In: Gretsi 2023 - XXIXème Colloque Francophone de Traitement du Signal et des Images. Grenoble, France, 2023, pp. 1–4. URL: https://hal.science/hal-04310386.

[38] L. Narmour, S. Derrien and S. Rajopadhye. 'Automatic Algorithm-Based Fault Tolerance (AABFT) of Stencil Computations'. In: *PACT '23: Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*. PACT 2023 - International Conference on Parallel Architectures and Compilation Techniques. Vienna, Austria, 2023, pp. 1–12. URL: https://inria.hal.science/hal-04394874.

[39] P. R. Nikiema, A. Kritikakou, M. Traiola and O. Sentieys. 'Impact of Transient Faults on Timing Behavior and Mitigation with Near-Zero WCET Overhead'. In: ECRTS 2023 - 35th Euromicro Conference on Real-Time Systems. Vienna, Austria: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, pp. 1–22. DOI: 10.4230/LIPIcs.ECRTS.2023.15. URL: https://hal.science/hal-04397374.

[40] P. R. Nikiema, A. Kritikakou, M. Traiola, O. Sentieys and O. Sentieys. 'Design with low complexity fine-grained Dual Core Lock-Step (DCLS) RISC-V processors'. In: DSN 2023 - 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Porto, Portugal: IEEE, 2023, pp. 224–229. DOI: 10.1109/DSN-S58398.2023.00062. URL: https://hal.science/hal-04397673.

[41] P. R. Nikiema, A. Palumbo, A. Aasma, L. Cassano, A. Kritikakou, A. Kulmala, J. Lukkarila, M. Ottavi, R. Psiakis and M. Traiola. 'Towards Dependable RISC-V Cores for Edge Computing Devices'. In: IOLTS 2023 - IEEE 29th International Symposium on On-Line Testing and Robust System Design. Crete, Greece: IEEE, 2023, pp. 1–7. DOI: 10.1109/IOLTS59296.2023.10224862. URL: https://hal.science/hal-04397384.

[42] J. Paturel, C. Quinson, M. Quinson and S. Rokicki. 'SmolPhone: a smartphone with energy limits'. In: IGSC 2023 - 14th International Green and Sustainable Computing. Toronto, Canada, 28th Oct. 2023, p. 4. URL: https://inria.hal.science/hal-04156447.

[43] A. Piri, S. Pappalardo, S. Barone, M. Barbareschi, B. Deveautour, M. Traiola, I. O'connor and A. Bosio. 'Input-aware accuracy characterization for approximate circuits'. In: DSN-W 2023 - 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops. Porto, Portugal: IEEE, 2023, pp. 179–182. DOI: 10.1109/DSN-W58399.2023.00050. URL: https://inria.hal.science/hal-04399159.

[44] B. Rossigneux, I. Kucher, V. Lorrain and E. Casseau. 'Surround the Nonlinearity: Inserting Foldable Convolutional Autoencoders to Reduce Activation Footprint'. In: *proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops*. ICCVW 2023 - IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE, 2023, pp. 1399–1403. DOI: 10.1109/ICCVW60793.2023.00152. URL: https://inria.hal.science/hal-04400777.

[45] M. Traiola, A. Kritikakou and O. Sentieys. 'A machine-learning-guided framework for fault-tolerant DNNs'. In: DATE 2023 – 26th IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium, 17th Apr. 2023, pp. 1–2. URL: https://inria.hal.science/hal-03887681.

[46] M. Traiola, A. Kritikakou and O. Sentieys. 'harDNNing: a machine-learning-based framework for fault tolerance assessment and protection of DNNs'. In: ETS 2023 - IEEE European Test Symposium. Venise, Italy: IEEE, May 2023, pp. 1–6. URL: https://hal.science/hal-04087375.

Conferences without proceedings

[47] F. Fernandes dos Santos, A. Kritikakou and O. Sentieys. 'Reliability evaluation of Convolutional Neural Network's basic operations on a RISC-V processor'. In: NSREC 2023 - IEEE Nuclear & Space Radiation Effects Conference. Kansas City, MO, United States: IEEE, 24th July 2023, pp. 1–6. URL: https://inria.hal.science/hal-04047058.

[48] F. Fernandes dos Santos, P. Rech, A. Kritikakou and O. Sentieys. 'Neutron-Induced Error Rate of Vision Transformer Models on GPUs'. In: RADECS - RADiation and its Effects on Components and Systems Conference. Toulouse, France, 25th Sept. 2023. URL: https://inria.hal.science/hal-04124814.

[49] J.-D. Guerrero-Balaguera, J. E. Rodriguez Condia, F. F. dos Santos, M. Sonza and P. Rech. 'Understanding the Effects of Permanent Faults in GPU's Parallelism Management and Control Units'. In: SC 2023 - ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis. Denver, United States: IEEE, 12th Nov. 2023, pp. 1–12. URL: https://inria.hal.science/hal-04132590.

[50] M. Lemaire, D. Massicotte, J. Poupart, P. Quinton and S. Rajopadhye. 'Polyhedra at Work: Automatic Generation of VHDL Code for the Sherman-Morrison Formula'. In: Impact24. Munich, Germany, 17th Jan. 2024. URL: https://inria.hal.science/hal-04401934.

[51] M. Traiola, F. Fernandes Dos Santos, O. Sentieys and A. Kritikakou. 'Impact of High-Level-Synthesis on Reliability of Neural Network Hardware Accelerators'. In: NSREC 2023 - IEEE Nuclear & Space Radiation Effects Conference. Kansas City (US), United States, 24th July 2023, pp. 1–5. URL: https://inria.hal.science/hal-04113282.

Scientific book chapters

[52] M. Barbareschi, S. Barone, A. Bosio and M. Traiola. 'Automatic Approximation of Computer Systems Through Multi-objective Optimization'. In: *Design and Applications of Emerging Computer Systems*. Springer Nature Switzerland, 17th Aug. 2024, pp. 383–420. DOI: 10.1007/978-3-031-42478-6_15. URL: https://inria.hal.science/hal-04396685.

Doctoral dissertations and habilitation theses

[53] T. Allenet. 'Quantization and adversarial robustness of embedded deep neural networks'. Université de Rennes, 24th Mar. 2023. URL: https://theses.hal.science/tel-04136202.

[54] M. T. Cong. 'Hardware accelerated simulation and automatic design of heterogeneous architecture'. Université de Rennes, 15th Mar. 2023. URL: https://theses.hal.science/tel-04136213.

[55] V.-P. Ha. 'Contributions to the scalability of automatic precision tuning'. Université de Rennes, 10th Mar. 2023. URL: https://theses.hal.science/tel-04189422.

11.3 Cited publications

[56] 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.

[57] 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.

[58] 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.

[59] D. Deb, R. M.K. and J. Jose. 'FlitZip: Effective Packet Compression for NoC in MultiProcessor System-on-Chip'. In: *IEEE Transactions on Parallel and Distributed Systems* 33.1 (2022), pp. 117–128. DOI: 10.1109/TPDS.2021.3090315.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

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

[66] 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.

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

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

[69] 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.

[70] 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.