UMR IRISA

Activity Report 2022

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, UNIV RENNES I, Advanced Research Position, HDR]
- François Charot [INRIA, Researcher]
- Silviu-Ioan Filip [INRIA, Researcher]
- Patrice Quinton [ENS RENNES, Emeritus]
- Marcello Traiola [INRIA, Researcher]

Faculty Members

- Olivier Sentieys [Team leader, UNIV RENNES I, Professor, HDR]
- Emmanuel Casseau [UNIV RENNES I, Professor, HDR]
- Daniel Chillet [UNIV RENNES I, Professor, HDR]
- Steven Derrien [UNIV RENNES I, Professor, HDR]
- Cédric Killian [UNIV RENNES I, Associate Professor]
- Angeliki Kritikakou [UNIV RENNES I, Associate Professor]
- Simon Rokicki [ENS RENNES, Associate Professor]

Post-Doctoral Fellows

- Sonia Barrios Pereira [INRIA]
- Abhijit Das [UNIV RENNES I]
- Fernando Fernandes Dos Santos [INRIA]

PhD Students

- Hamza Amara [UNIV RENNES I, from Oct 2022]
- Herinomena Andrianatrehina [INRIA, from Nov 2022]
- Gaetan Barret [ORANGE, CIFRE, from Oct 2022]
- Sami Ben Ali [INRIA]
- Benoit Coqueret [Thales, CIFRE, from Nov 2022]
- Leo De La Fuente [CEA, Grenoble]
- Paul Estano [INRIA, from Feb 2022]
- Corentin Ferry [UNIV RENNES I]
- Cedric Gernigon [INRIA]
- Jean-Michel Gorius [UNIV RENNES I]
- Wilfread Guilleme [UNIV RENNES I, from Oct 2022]
- Ibrahim Krayem [UNIV RENNES I]

- Seungah Lee [UNIV RENNES I]
- Guillaume Lomet [INRIA, from Oct 2022]
- Amélie Marotta [INRIA]
- Romain Mercier [INRIA, until Feb 2022]
- Louis Narmour [UNIV RENNES I]
- Pegdwende Nikiema [UNIV RENNES I, from Sep 2022]
- Léo Pradels [SAFRAN]
- Baptiste Rossigneux [CEA, from Nov 2022, Saclay]
- Louis Savary [UNIV RENNES I, from Sep 2022]

Technical Staff

- Ludovic Claudepierre [UNIV RENNES I, Engineer, from Nov 2022]
- Pierre Halle [INRIA, Engineer, until Apr 2022]
- Romain Mercier [UNIV RENNES I, Engineer, from Mar 2022]
- Arash Nejat [INRIA, Engineer]
- Joseph Paturel [INRIA, Engineer, from Apr 2022]
- Dikshanya Lashmi Ramaswamy [INRIA, Engineer, from Sep 2022]

Administrative Assistants

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

2 Overall objectives

Energy efficiency has now become one of the main requirements for virtually all computing platforms [78]. 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 [77, 76], 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 [73]. The main driver today is energy efficiency—small embedded devices need specialized hardware to operate under power/energy constraints. In the next ten years, we expect specializations to become even more common to meet increasing demands for performance. In particular, high-throughput workloads traditionally ran on servers (e.g., computational science and machine learning) will offload (parts of) their computations to accelerators. We are already seeing some instances of such specialization, most notably accelerators for neural networks that use clusters of nodes equipped with FPGAs and/or ASICs.

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

2.2 Design Stack for Custom Hardware

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

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

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

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

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

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

2.3 Objectives of TARAN: Facilitating Cross-Layer Optimization

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

3 Research program

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

- **Accelerators**: Hardware accelerators will become more and more common, and we must develop techniques to make accelerator design more accessible. The important challenge is raising the level of abstraction without sacrificing performance. However, higher level of abstraction coupled with domain-specific knowledge is also a great opportunity to widen the scope of accelerators.
- **Accurate Computing**: Most computing today is performed with significant over-provisioning of output quality or precision. Carefully selecting the various parameters, ranging from algorithms to arithmetic, to compute with just the right quality is necessary for further efficiency. Such fine tuning of elements affecting application quality is extremely time consuming and requires domain knowledge to be fully utilized.
- **Resilient Computing**: As we approach the limit of CMOS scaling, it becomes increasingly unlikely for a computing device to be fully functional due to various sources of faults. Thus, techniques to maintain efficiency in the presence of faults will be important. Generally applicable techniques, such as replication, come with significant overheads. Developing techniques tailored to each application will be necessary for computing contexts where reliability is critical.
- **Embracing Emerging Technologies**: Certain computing platforms, such as ultra-low power devices and embedded many-cores, have specific design constraints that make traditional components unfit. However, emerging technologies such as Non-Volatile Memory and Silicon Photonics cannot simply be used as a substitute. Effectively integrating more recent technologies is an important challenge for these specialized computing platforms.

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

3.1 Accelerators

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

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

Positioning Hardware acceleration has already proved its efficiency in many datacenter, cloud-computing or embedded high-performance computing (HPC) applications: machine learning, web search, data mining, database access, information security, cryptography, financial, image/signal/video processing, etc. For example, the work at Microsoft in accelerating the Bing web search engine with large-scale reconfigurable fabrics has shown to improve the ranking throughput of each server by 95% [72], and the increasing need for acceleration of deep learning workloads [84].

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 [74, 80]. On this topic, we have been pursuing our efforts to propose tools whose design principles are based on a tight coupling between the compiler and the target hardware architectures.

3.2 Accurate Computing

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

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

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

The need for further improvements in energy efficiency has led to renewed interest in approximation techniques in the recent years [81]. 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 [75].

3.3 Resilient Computing

Key Investigators: E. Casseau, D. Chillet, C. Killian, A. Kritikakou, O. Sentieys.
With advanced technology nodes and the emergence of new devices pressured by the end of Moore's law, manufacturing problems and process variations strongly influence electrical parameters of circuits and architectures [79], leading to dramatically reduced yield rates [82]. Transient errors caused by particles or radiations will also more and more often occur during execution [85, 83], and process variability will prevent predicting chip performance (e.g., frequency, power, leakage) without a self-characterization at run time. On the other hand, many systems are under constant attacks from intruders and security has become of utmost importance.

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

3.4 Embracing Emerging Technologies

Key Investigators: D. Chillet, S. Derrien, C. Killian, O. Sentieys.
Domain specific accelerators have more exploratory freedom to take advantage of non-conventional technologies that are too specialized for general purpose use. Examples of such technologies include optical interconnects for Network-on-Chip (NoC) and Non-Volatile Memory (NVM) for low-power sensor nodes. The objective of this research direction is to explore the use of such technologies, and find appropriate application domains. The primary cross-layer interaction is expected from Hardware Design

to accommodate non-conventional Technologies. However, this research direction may also involve Runtime and Compilers.

4 Application domains

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

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

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

5 New software and platforms

5.1 New software

5.1.1 Gecos

Name: Generic Compiler Suite

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

Scientific Description: The Gecos (Generic Compiler Suite) project is a source-to-source compiler infrastructure 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: This year, we have proposed a fully automated hardware synthesis flow based on a source-to-source compiler that identifies and explores intricate speculation configurations to generate speculative hardware accelerators [25].

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

Publication: hal-03714101

Contact: Steven Derrien

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

Partner: Université de Rennes 1

5.1.2 SmartSense

Name: Sensor-Aided Non-Intrusive Load Monitoring

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

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

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

Contact: Olivier Sentieys

5.1.3 TypEx

Name: Type Exploration Tool

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

Scientific Description: The main goal of TypEx is to explore the design space spanned by possible number formats in the context of High-Level Synthesis. TypEx takes a C code written using floating-point datatypes specifying the application to be explored. The tool also takes as inputs a cost model as well as some user constraints and generates a C code where the floating-point datatypes are replaced by the wordlengths found after exploration. The best set of wordlengths is the one found by the tool that respects the accuracy constraint given and that minimizes a parametrized cost function.

Functional Description: TypEx is a tool designed to automatically determine custom number representations and word-lengths (i.e., bit-width) for FPGAs and ASIC designs at the C source level. TypEx is available open-source at https://gitlab.inria.fr/gecos/gecos-float2fix. See README.md for detailed instructions on how to install the software.

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

Contact: Olivier Sentieys

5.2 New platforms

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

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.

- Participants: Silviu-Ioan Filip
- Partners: Univ Rennes
- Contact: Silviu-Ioan Filip
- URL: https://github.com/mptorch/mptorch

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

KEYWORDS: function approximation, FPGA hardware implementation generator

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

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

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

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

KEYWORDS: FIR filter design, multiplierless hardware implementation generator

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

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

5.2.4 Hybrid-DBT

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

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

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

5.2.5 Comet

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

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

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

6 New results

6.1 Improving Memory Throughput of Hardware Accelerators

Participants: Steven Derrien, Corentin Ferry, Tomofumi Yuki.

Offloading compute-intensive kernels to hardware accelerators relies on the large degree of parallelism offered by these platforms. However, the effective bandwidth of the memory interface often causes a bottleneck, hindering the accelerator's effective performance. Techniques enabling data reuse, such as tiling, lower the pressure on memory traffic but still often leave the 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.) [24].

6.2 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 [5]. In particular, speculative loop pipelining appears as a promising approach as it can handle both control-flow and memory speculations within a classical HLS framework. Our last contribution in this topic consists in proposing a fully automated hardware synthesis flow based on a source-to-source compiler that identifies and explores intricate speculation configurations to generate speculative hardware accelerators [25]. We demonstrate that the proposed tool is capable of generating efficient accelerators for several real-life applications, which greatly benefit from the use of speculation.

6.3 Design Space Exploration for IoT Processors Platforms

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

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

We address this issue by providing a flow capable of automatically synthesizing pipelined micro-architectures directly from an Instruction Set Simulator in C/C++. Our flow is based on HLS technology and bridges part of the gap between Instruction Set Processor design flows and High-Level Synthesis tools by taking advantage of speculative loop pipelining. Our results show that our flow is general enough to support a variety of ISA and micro-architectural extensions, and is capable of producing circuits that are competitive with manually designed cores. The first results from this work have been presented at the IEEE FPT conference [41]. We are currently working to enable the synthesis of more complex micro-architecture features (e.g. predictors, in-order superscalar, etc.). We also plan to study how security issues could be handled in our proposed processor design flow.

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

Participants: Olivier Sentieys, Angeliki Kritikakou.

Heterogeneous multi-core platforms, such as ARM big.LITTLE are widely used to execute embedded applications under multiple and contradictory constraints, such as energy consumption and real-time execution. To fulfill these constraints and optimize system performance, application tasks should be efficiently mapped on multi-core platforms. Embedded applications are usually tolerant to approximated results but acceptable Quality-of-Service (QoS). Modeling embedded applications by using the elastic task model, namely, Imprecise Computation (IC) task model, can balance system QoS, energy consumption, and real-time performance during task deployment. However, state-of-the-art approaches seldom consider the problem of IC task deployment on heterogeneous multi-core platforms. They typically neglect task migration, which can improve the solutions due to its flexibility during the task deployment process. We proposed a novel QoS-aware task deployment method to maximize system QoS under energy and real-time constraints, where frequency assignment, task allocation, scheduling, and migration are optimized simultaneously [28]. The task deployment problem is formulated as mixed-integer non-linear programming. Then, it is linearized to mixed-integer linear programming to find the optimal solution. Furthermore, based on problem structure and problem decomposition, we propose a novel heuristic with low computational complexity. The sub-problems regarding frequency assignment, task allocation, scheduling, and adjustment are considered and solved in sequence. Finally, the simulation results show that the proposed task deployment method improves the system QoS by 31.2% on average (up to 112.8%) compared to the state-of-the-art methods and the designed heuristic achieves about 53.9% (on average) performance of the optimal solution with a negligible computing time. This work is done in collaboration with Lei Mo, School of Automation, Southeast University (China).

6.5 Training Deep Neural Networks with Low-Precision Accelerators

Participants: Silviu Filip, Olivier Sentieys.

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

The most compute-intensive stage of deep neural network (DNN) training is matrix multiplication where the multiply-accumulate (MAC) operator is key. To reduce training costs, in this work [53] we consider using low-precision arithmetic for MAC operations. While low-precision training has been investigated in prior work, the focus has been on reducing the number of bits in weights or activations without compromising accuracy. In contrast, the focus in this work is on implementation details beyond weight or activation width that affect area and accuracy. In particular, we investigate the impact of fixed- versus floating-point representations, multiplier rounding, and floating-point exceptional value support. Results suggest that (1) low-precision floating-point is more area-effective than fixed-point for multiplication, (2) standard IEEE-754 rules for subnormals, NaNs, and intermediate rounding serve little to no value in terms of accuracy but contribute significantly to area, (3) low-precision MACs require an adaptive loss-scaling step during training to compensate for limited representation range, and (4) fixed-point is more area-effective for accumulation, but the cost of format conversion and downstream logic can swamp the savings. Finally, we note that future work should investigate accumulation structures beyond the MAC level to achieve further gains.

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

We also published a book chapter that explores and reviews how Approximate Computing can improve the performance and energy efficiency of hardware accelerators in Deep Learning applications during inference and training [60].

6.6 Quantization-Aware Training for Efficient DNN Inference

Participants: Thibault Allenet, Olivier Sentieys.

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 [32], 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.

6.7 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 [42], 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%.

We also published a book chapter that reviews low-precision arithmetic operators and custom number representations [61]. This chapter is part of a book on Approximate Computing Techniques [58], Olivier Sentieys from Taran being one of the editors of this book.

6.8 A Genetic-algorithm-based Approach to the Design of Approximate DCT Hardware Accelerators

Participants: Marcello Traiola.

As modern applications demand an unprecedented level of computational resources, traditional computing system design paradigms are no longer adequate to guarantee significant performance enhancement at an affordable cost. Approximate Computing (AxC) has been introduced as a potential candidate to achieve better computational performance by relaxing non-critical functional system specifications. In [20], we propose a systematic and high-abstraction-level approach allowing the automatic generation of near Pareto-optimal approximate configurations for a Discrete Cosine Transform (DCT) hardware accelerator. We obtain the approximate variants by using approximate operations, having configurable approximation degree, rather than full-precise ones. We use a genetic searching algorithm to find the appropriate tuning of the approximation degree, leading to optimal trade-offs between accuracy and gains. Finally, to evaluate the actual HW gains, we synthesize non-dominated approximate DCT variants for two different target technologies, namely, Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs). Experimental results show that the proposed approach allows performing a meaningful exploration of the design space to find the best tradeoffs in a reasonable time. Indeed, compared to the state-of-the-art work on approximate DCT, the proposed approach allows an 18% average energy improvement while providing at the same time image quality improvement.

6.9 Design of Finite Impulse Response Filters with Minimal Number of Adders

Participants: Silviu Filip.

This work [26] presents two novel methods that simultaneously optimize both the design of a finite impulse response (FIR) filter and its multiplierless hardware implementation. We use integer linear programming (ILP) to minimize the number of adders used to implement a direct/transposed FIR filter adhering to a given frequency specification. The proposed algorithms work by either fixing the number

of adders used to implement the products (multiplier block adders) or by bounding the adder depth (AD) used for these products. The latter can be used to design filters with minimal AD for low power applications. In contrast to previous multiplierless FIR filter approaches, the methods introduced here ensure adder count optimality. We perform extensive numerical experiments which demonstrate that our simultaneous filter design approach yields results which are in many cases on par or better than those in the literature.

6.10 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 [36]. 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.

6.11 Assertion-Aware Approximate Computing Design Exploration on Behavioral Models

Participants: Marcello Traiola.

Several design exploration approaches and metrics have been proposed so far to identify the approximation targets, but only a few of them exploit information derived from assertion-based verification (ABV). To fill in the gap, in [35] we propose an ABV-based methodology to guide the AxC design exploration of behavioral descriptions. Assertions are automatically mined from the simulation traces of the original design to capture the golden behaviours. Then, we define a metric to predict the impact of approximating model statements on the design accuracy. The metric is computed by a function based on the syntax tree of the mined assertions, their support, and the information derived from the variable dependency graph of the design. It is therefore used, together with functional coverage information, in a sorting procedure for AxC design space exploration to select the target statements for approximation.

6.12 Input-Aware Approximate Computing

Participants: Marcello Traiola.

An important amount of work has been done in proposing approximate versions of basic operations, using fewer resources. From a hardware standpoint, several approximate arithmetic operations have been proposed. Although effective, such approximate hardware operators are not tailored to a specific final application. Thus, their effectiveness will depend on the actual application using them. Taking into account the target application and the related input data distribution, the final energy efficiency can be pushed further. In [49], we showcase the advantage of considering the data distribution by designing

an input-aware approximate multiplier specifically intended for a high pass FIR filter, where the input distribution pattern for one operand is not uniform. Experimental results show that we can significantly reduce the power consumption while keeping an error rate lower than state of the art approximate multipliers.

6.13 Test and Reliability of Approximate Hardware

Participants: Marcello Traiola.

The undeniable need of energy efficiency in today's devices is leading to the adoption of innovative computing paradigms—such as Approximate Computing. As this paradigm is gaining increasing interest, important challenges, as well as opportunities, arise concerning the dependability of those systems. The book chapter [62] focuses on test and reliability issues related to approximate hardware systems. It covers problems and solutions concerning the impact of the approximation on hardware defect classification, test generation, and test application. Moreover, the impact of the approximation on the fault tolerance is discussed, along with related design solutions to mitigate it.

6.14 Design, Verification, Test, and In-Field Implications of Approximate Digital Integrated Circuits

Participants: Marcello Traiola.

While Approximate Computing allows many improvements when looking at systems' performance, energy efficiency, and complexity, it poses significant challenges regarding the design, the verification, the test, and the in-field reliability of Approximate Digital Integrated Circuits. This chapter [59] covers these aspects, leveraging the authors' experience in the field to present state-of-the-art solutions to apply during the different development phases of an Approximate Computing system.

6.15 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. Here, 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.

6.16 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 and GPU-based platforms. Although RISC-V architectures have gained importance in the last years, the application's error rate on RISC-V processors is not significantly evaluated, as it has been done for standard x86 processors. To address this limitation, we investigate the error rate of a commercial RISC-V ASIC platform, the GAP8, exposed to a neutron beam. The results show that for computing-intensive applications, such as classification Convolutional Neural Networks (CNN), the error rate can be 3.2x higher than the average error rate. Most (96.12%) of the errors on the CNN do not generate misclassifications, while the major source of incorrect interruptions is application hangs [39, 70].

Graphics Processing Units (GPUs) can execute floating-point operations with mixed-precisions, such as INT8, FP16, Bfloat, FP32, and FP64, and include hardaware dedicated to multiplication, e.g., NVIDIA GPUs have tensor cores that perform 4x4 FP16 matrix multiplication is performed in a single instruction. We have measured the error rate of mixed precision algorithms related to DNNs by exposing an NVIDIA Volta GPU (Tesla V100) to a beam of neutrons [52], considering multiple floating-point precisions of a General Matrix Multiplication (GEMM), such as FP16, FP32, FP64, and tensor cores using FP16. Our results show that the smaller the precision, the smaller the execution time and error rate, i.e., the error rate of FP16, FP32, and FP64 GEMM are 1.6x, 2.5x, and 3.3x higher than the version that uses FP16 and Tensor Cores units, respectively.

Simulation-based system reliability analysis is performed by injecting faults at different abstraction layers of the system. Existing approaches either partially model the studied system's fault masking capabilities, losing accuracy, or require prohibitive estimation times. To deal with this limitation, we propose a vulnerability analysis approach that combines gate-level fault injection with microarchitecture-level Cycle-Accurate and Bit-Accurate simulation, achieving low estimation time [45]. Faults both in sequential and combinational logic are considered and fault masking is modeled at gate-level, microarchitecture-level and application-level, maintaining accuracy. The approach highlights that a significant number of Multiple-Event-Upsets (MEUs) are derived by faults (SETs) in the combinational logic. These MEUs can be significantly large in size and they not disturb only adjacent bits. Thus, radiation- induced faults should not be modeled only with SEU, but also with (significantly large) MEUs. Our case-study is a RISC-V processor. Obtained results show a more than 8% reduction in masked errors, increasing system failures by more than 55% compared to standard fault injection approaches, which fully validates the hypothesis on the impact of MEUs to the vulnerability and our analysis flow. The above approach has been enhanced in order to expose the timing impact of transient faults, which can affect the temporal correctness of the system [44].

Hybrid approaches have been proposed for the reliability evaluation of DNN executed on GPUs, since the hardware architecture is highly complex and the software frameworks are composed of many layers of abstraction. While software-level fault injection is a common and fast way to evaluate the reliability of complex applications, it may produce unrealistic results as it has limited access to the hardware resources, and the adopted fault models may be too naive (i.e., single and double-bit flip). Contrarily, physical fault injection with a neutron beam provides realistic error rates, but lacks fault propagation visibility. Thus, we proposed a DNN fault model characterization, combining neutron beam experiments and fault injection at the software level [57, 23]. GPUs running General Matrix Multiplication (GEMM) and DNNs are exposed to beam neutrons to measure their error rate. On DNNs, we observe that the percentage of critical errors can be up to 61%, showing that ECC is ineffective in reducing critical errors. We then performed a complementary software-level fault injection using fault models derived from RTL simulations. Our results show that by injecting a cocktail of complex fault models, the YOLOv3 misdetection rate is validated to be very close to the rate measured with beam experiments, which is

8.66x higher than the one measured with fault injection using only naive single-bit flips.

6.17 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 [54], 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.

6.18 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 [31] 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.

6.19 Improving the Fault Resilience of Neural Network Applications Through Security Mechanisms

Participants: Marcello Traiola.

Numerous electronic systems store valuable intellectual property (IP) information inside non-volatile memories. In order to protect the integrity of such sensitive information from an unauthorized access or modification, encryption mechanisms are employed. From a reliability standpoint, such information can be vital to the system's functionality and thus, dedicated techniques are employed to detect possible reliability threats (e.g., transient faults in the memory content). In [38], we explore the capability

of encryption mechanisms to guarantee protection from both unauthorized access and faults, while considering a Convolutional Neural Network application whose weights represent the valuable IP of the system. Experimental results show that it is possible to achieve very high fault detection rates, thus exploiting the benefits of security mechanisms for reliability purposes as well.

6.20 Selective Hardening of Critical Neurons in Deep Neural Networks

Participants: Marcello Traiola.

In the literature, it is argued that Deep Neural Networks (DNNs) possess a certain degree of robustness mainly for two reasons: their distributed and parallel architecture, and their redundancy introduced due to over provisioning. Indeed, they are made, as a matter of fact, of more neurons with respect to the minimal number required to perform the computations. It means that they could withstand errors in a bounded number of neurons and continue to function properly. However, it is also known that different neurons in DNNs have divergent fault tolerance capabilities. Neurons that contribute the least to the final prediction accuracy are less sensitive to errors. Conversely, the neurons that contribute most are considered critical because errors within them could seriously compromise the correct functionality of the DNN. The paper in [51] presents a software methodology based on a Triple Modular Redundancy technique, which aims at improving the overall reliability of the DNN, by selectively protecting a reduced set of critical neurons. Our findings indicate that the robustness of the DNNs can be enhanced, clearly, at the cost of a larger memory footprint and a small increase in the total execution time. The trade-offs as well as the improvements are discussed in the work by exploiting two DNN architectures: ResNet and DenseNet trained and tested on CIFAR-10.

6.21 Fault-Tolerant Microarchitectures

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

Instruction Level Parallelism (ILP) of applications is typically limited and variant in time, thus during application execution some processor Function Units (FUs) may not be used all the time. Therefore, these idle FUs can be used to execute replicated instructions, improving reliability. However, existing approaches either schedule the execution of replicated instructions based on compiler schedule or consider processors with identical FUs, able to execute any instruction type. The former approach has a negative impact on performance, whereas the later approach is not applicable on processors with heterogeneous FUs. We propose a hardware mechanism for processors with heterogeneous FUs that dynamically replicates instructions and schedules both original and replicated instructions considering space and time scheduling [29]. The proposed approach uses a small scheduling window of two cycles, leading to a hardware mechanism with small hardware area. In order to perform such a flexible dynamic instruction scheduling, switches are required, which, however, increase the hardware area. To reduce the area overhead, a cluster-based approach is proposed, enabling scalability for larger hardware designs. The proposed mechanism is implemented on VEX VLIW processor. The obtained results show an average speed-up of 24.99% in performance with an almost 10% area and power overhead, when time scheduling is also considered on top of space scheduling.

Recent findings indicate that transient hardware faults may corrupt the DNN models prediction dramatically. For instance, the radiation-induced misprediction probability can be so high to impede a safe deployment of DNNs models at scale, urging the need for efficient and effective hardening solutions. In this work, we propose to tackle the reliability issue both at training and model design time [69]. First, we show that vanilla models are highly affected by transient faults, that can induce performances to drop up to 37%. Hence, we provide three zero-overhead solutions, based on DNN re-design and re-train, that can improve DNNs reliability to transient faults up to one order of magnitude. We complement our work with extensive ablation studies to quantify the gain in performance of each hardening component. Our

results show that the proposed DNNs for image classification can be 10x (CIFAR 10 dataset) and 3.2x (CIFAR 100 dataset) more reliable than the baseline DNN without protection.

We have proposed an effective methodology to identify the architectural vulnerable sites in GPUs modules, i.e. the locations that, if corrupted, most affect the correct instructions execution [30]. We first identify, through an innovative method based on Register-Transfer Level (RTL) fault injection experiments, the architectural vulnerabilities of a GPU model. Then, we mitigate the fault impact via selective hardening applied to the flip-flops that have been identified as critical. We evaluate three hardening strategies: Triple Modular Redundancy (TMR), Triple Modular Redundancy against SETs, and Dual Interlocked Storage Cells. The results gathered on a publicly available GPU Model (FlexGripPlus) considering functional units, pipeline registers, and warp scheduler controller show that our method can tolerate from 85% to 99% of faults in the pipeline registers, from 50% to 100% of faults in the functional units and up to 10% of faults in the warp scheduler, with a reduced hardware overhead (in the range of 58% to 94% when compared with traditional TMR).

6.22 Fault-Tolerant Networks-on-Chip

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

Network-on-Chip has become the main interconnect in the multicore/manycore era since the beginning of this decade. However, these systems become more sensitive to faults due to transistor shrinking size. In parallel, approximate computing appears as a new computation model for applications since several years. The main characteristic of these applications is to support the approximation of data, both for computations and for communications. To exploit this specific application property, we develop a fault-tolerant NoC to reduce the impact of faults on the data communications. To address this problem, we consider multiple permanent faults on router which cannot be managed by Error-Correcting Codes (ECCs), or at a high hardware cost. For that, we propose a bit-shuffling method to reduce the impact of faults on Most Significant Bits (MSBs), hence permanent faults only impact Least Significant Bits (LSBs) instead of MSBs reducing the errors impact.

To decrease hardware costs, we proposed a region-based bit-shuffling technique in [56], applied at a coarse-grain level, that trades off fault mitigation efficiency in order to save hardware costs. The obtained results show that the area and power overheads can be reduced from 48% to 33% and from 34% to 22%, respectively, with a small impact on the MSE [46].

Finally, we proposed a routing technique for the management of faulty paths in an NoC [43]. For non-critical data to transfer, the technique exploits the bit shuffling error mitigation method to propagate the data through faulty paths. For critical data, two methods are evaluated. The first one circumvents the faulty path by sending the data packet through a new faulty free path. This decision can increase the congestion in specific NoC area. To limit the congestion problem, the second method is based on flit duplication to transmit the packet through the faulty path. We show that this technique ensures the communications on a faulty NoC with a limit latency overhead compared to solutions using only routing adaptations [43].

6.23 Fault-Tolerant Task Deployment onto Multicore Systems

Participants: Emmanuel Casseau, Minyu Cui, Angeliki Kritikakou.

Task deployment plays an important role in the overall system performance, especially for complex architectures, since it affects not only the energy consumption but also the real-time response and reliability of the system. We are focusing on how to map and schedule tasks onto homogeneous processors under faults at design time. Dynamic Voltage/Frequency Scaling (DVFS) is typically used for energy saving, but with a negative impact on reliability, especially when the frequency is low. Using high frequencies to meet reliability and real-time constraints leads to high energy consumption, while multiple replicas at

lower frequencies may increase energy consumption. To 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 [21]. 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 [22]. To cope with the complexity of such problem, we have proposed mapping heuristics for task-level DVFS in order to reduce the time required to find a solution and thus enhance the scalability of the proposed approach [37].

Furthermore, in [47] a task deployment approach is proposed for multicore architectures with homogeneous cores connected with Network-on-Chip (NoC). The goal is to optimize the overall system energy consumption, including computation of the cores and communication of the NoC, under task reliability and real-time constraints. More precisely, the task deployment approach combines task allocation and scheduling, frequency assignment, task duplication, and multipath data routing. The task deployment problem is formulated using mixed-integer non-linear programming. To find the optimal solution, the original problem is equivalently transformed to mixed-integer linear programming, and solved by state-of-the-art solvers. Furthermore, a decomposition-based heuristic, with low computational complexity, is proposed to deal with scalability. This work is done in collaboration with Lei Mo School of Automation, Southeast University (China).

6.24 Optical Network-on-Chip for error resilient applications

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

The energy consumption of manycore is dominated by data transfers, which calls for energy- efficient and high-bandwidth interconnects. Classical electrical NoC solutions suffer from low scalability and low performance when the number of cores to connect becomes high. To tackle this challenge, integrated optics appears as promising technology to overcome the bandwidth limitations of electrical interconnects. However, this technology suffers from high power overhead related to low efficiency lasers. From these observations, the concept of approximate communications appears as interesting technique to reduce the power of lasers.

In this context, we have developed an approximate communication model for data exchanges based on laser power management. The data to transfer are classified into sensitive data and data which can be approximated without too much Quality of Service (QoS) degradation. From this classification, we are able to reduce the energy of communication by reducing the laser power of LSB bits (Least Significant Bits) and/or by truncating them, while the MSB bits are sent at nominal power level. The SNR of the LSBs is then reduced or truncated impacting the communication QoS. Furthermore, we also defined a distance-aware technique which takes account of both the communication distance and the quality of service to compute the laser power [27]. From these contributions, we have developed a simulation platform, based on Sniper, and we show that our solution is scalable and leads to 10% reduction in the total energy consumption, 35× reduction in the laser driver size, and 10× reduction in the laser controller compared to state-of-the-art solutions.

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

Participants: Joel Ortiz Sosa, Cédric Killian.

A key challenge for the deployment of nanophotonic interconnects is their high static power, which is induced by signal losses and devices calibration. To tackle this challenge, we propose to use Phase Change Material (PCM) to configure optical paths between writers and readers. The non-volatility of PCM elements and the high contrast between crystalline and amorphous phase states allow to bypass

unused readers, thus reducing losses and calibration requirements. We evaluate the efficiency of the proposed PCM-based interconnects using system level simulations carried out with SNIPER manycore simulator. For this purpose, we have modified the simulator to partition clusters according to executed applications. Simulation results show that bypassing readers using PCM leads up to 52% communication power saving [55].

6.26 A Design Space Exploration Framework for Memristor-Based Crossbar Architecture

Participants: Marcello Traiola.

In the literature, there are few studies describing how to implement Boolean logic functions as a memristor-based crossbar architecture and some solutions have been actually proposed targeting back-end synthesis. However, there is a lack of methodologies and tools for the synthesis automation. The main goal of [33] is to perform a Design Space Exploration (DSE) in order to analyze and compare the impact of the most used optimization algorithms on a memristor-based crossbar architecture. The results carried out on 102 circuits lead us to identify the best optimization approach, in terms of area/energy/delay. The presented results can also be considered as a reference (benchmarking) for comparing future work

7 Bilateral contracts and grants with industry

7.1 Bilateral contracts with industry

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.

7.2 Bilateral Grants with Industry

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

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

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

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.

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.

7.3 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).

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 Inria Associate Team

IntelliVIS

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

Duration: 2019 - 2022

Coordinator: Olivier Sentieys

Partners: IIT Goa (India)

Inria contact: Olivier Sentieys

Summary: The proposed collaborative research work is focused on the design and development of artificial intelligence based embedded vision architectures for cyber physical systems (CPS) and edge devices.

EdgeTrain

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)

Inria contact: Silviu-Ioan Filip

Other people involved: Olivier Sentieys, Guy Lemieux (UBC)

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.

8.1.2 Inria International Partners

LRS

Title: Loop unRolling Stones: compiling in the polyhedral model

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

Inria contact: Steven Derrien

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

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

8.2 International research visitors

8.2.1 Visits of international scientists

Louis Narmour 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, is visiting TARAN from Nov. 2020 for two years.

8.2.2 Visits to international teams

Corentin Ferry 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.

8.3 National initiatives

8.3.1 ANR AdequateDL

Participants: Olivier Sentieys, Silviu-Ioan Filip.

- Program: ANR PRC
- Project acronym: AdequateDL
- Project title: Approximating Deep Learning Accelerators
- Duration: Jan. 2019 - Dec. 2023
- Coordinator: TARAN
- Other partners: INL, LIRMM, CEA-LIST

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

8.3.2 ANR RAKES

Participants: Olivier Sentieys, Cédric Killian, 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 2023
- Coordinator: TIMA
- Other partners: TIMA, TARAN, Lab-STICC

The efficient exploitation by software developers of multi/many-core architectures is tricky, especially when the specificities of the machine are visible to the application software. To limit the dependencies to the architecture, the generally accepted vision of the parallelism assumes a coherent shared memory and a few, either point to point or collective, synchronization primitives. However, because of the difference of speed between the processors and the main memory, fast and small dedicated hardware controlled memories containing copies of parts of the main memory (a.k.a caches) are used. Keeping these distributed copies up-to-date and synchronizing the accesses to shared data, requires to distribute and share information between some if not all the nodes. By nature, radio communications provide broadcast capabilities at negligible latency, they have thus the potential to disseminate information very

quickly at the scale of a circuit and thus to be an opening for solving these issues. In the RAKES project, we intend to study how wireless communications can solve the scalability of the abovementioned problems, by using mixed wired/wireless Network on Chip. We plan to study several alternatives and to provide (a) a virtual platform for evaluation of the solutions and (b) an actual implementation of the solutions.

8.3.3 ANR Opticall2

Participants: Olivier Sentieys, Cédric Killian, Daniel Chillet.

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

The aim of Opticall2 is to design broadcast-enabled optical communication links in manycore architectures at wavelengths around 1.3um. We aim to fabricate an optical broadcast link for which the optical power is equally shared by all the destinations using design techniques (different diode absorption lengths, trade-off depending on the current point in the circuit and the insertion losses). No optical switches will be used, which will allow the link latency to be minimized and will lead to deterministic communication times, which are both key features for efficient cache coherence protocols. The second main objective of Opticall2 is to propose and design a new broadcast-aware cache coherence communication protocol allowing hundreds of computing clusters and memories to be interconnected, which is well adapted to the broadcast-enabled optical communication links. We expect better performance for the parallel execution of benchmark programs, and lower overall power consumption, specifically that due to invalidation or update messages.

8.3.4 ANR SHNOC

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

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

The goal of the SHNoC project is to tackle one of the manycore interconnect issues (scalability in terms of energy consumption and latency provided by the communication medium) by mixing emerging technologies. Technology evolution has allowed for the integration of silicon photonics and wireless on-chip communications, creating Optical and Wireless NoCs (ONoCs and WNoCs, respectively) paradigms. The recent publications highlight advantages and drawbacks for each technology: WNoCs are efficient for broadcast, ONoCs have low latency and high integrated density (throughput/sqcm) but are inefficient in multicast, while ENoCs are still the most efficient solution for small/average NoC size. The first

contribution of this project is to propose a fast exploration methodology based on analytical models of the hybrid NoC instead of using time consuming manycore simulators. This will allow exploration to determine the number of antennas for the WNoC, the amount of embedded lasers sources for the ONoC and the routers architecture for the ENoC. The second main contribution is to provide quality of service of communication by determining, at run-time, the best path among the three NoCs with respect to a target, e.g. minimizing the latency or energy. We expect to demonstrate that the three technologies are more efficient when jointly used and combined, with respect to traffic characteristics between cores and quality of service targeted.

8.3.5 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.

8.3.6 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.

8.3.7 DGA/INRIA Sniffer

Participants: Olivier Sentieys.

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

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

8.3.8 Labex CominLabs - LeanAI (2021-2024)

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

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

9 Dissemination

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

Member of the organizing committees

- D. Chillet was in the Organizing Committee of HiPEAC Rapido'22.
- A. Kritikakou was in the Organizing Committee of INRIA Scientific days (JSI) 2022.
- M. Traiola was in the Organizing Committees of IEEE/ACM Design Automation and Test in Europe (DATE) 2022, IEEE VLSI Test Symposium (VTS) 2022, IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS) 2022, IEEE International Workshop on Silicon Lifecycle Management (SLM) 2022, European Automotive Reliability, Test, and Safety workshop (eARTS) 2022, and European Workshop on Silicon Lifecycle Management (eSLM) 2022

9.1.2 Scientific events: selection

Chair of conference program committees

- O. Sentieys was the Chair of the D9 Track on Architectural and Microarchitectural Design at IEEE/ACM DATE 2022.
- O. Sentieys served as a committee member in the IEEE EDAA Outstanding Dissertations Award (ODA) 2022.
- D. Chillet was the Session Program Co-Chair of Gretsi 2022.
- A. Kritikakou was the Chair for the Computer-Aided Design and Verification Track and the Student Research Forum at ISVLSI 2022.
- A. Kritikakou was the Chair for the Brief Presentations Track at RTAS 2022.
- A. Kritikakou was the Chair for the Artifact Evaluation at ECRTS 2022.
- M. Traiola was the Technical program Co-Chair for the IEEE Workshop on Silicon Errors in Logic – System Effects (SELSE) 2022

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. Derrien was a member of technical program committee of IEEE FPL, IEEE ASAP, IEEE/ACM PACT, ARC.
- A. Kritikakou was a member of technical program committee of IEEE RTSS, ECRTS, IEEE/ACM DATE, ISVLSI, SAMOS, DS-RT, RTNS, ARC, CPSCom, ComPAS.
- O. Sentieys was a member of technical program committee of IEEE/ACM DATE, IEEE FPL, ACM ENSSys, ACM SBCCI, IEEE ReConFig, FDL, ARC.
- C. Killian was a member of technical program committee of ACM NOCS, DATE, NOCARC.
- S. Rokicki was a member of technical program committee of CASES
- M. Traiola was a member of technical program committee of IEEE VTS, IEEE IOLTS, IEEE/ITRI VLSI-DAT, eARTS.

9.1.3 Journal

Member of the editorial boards

- D. Chillet is member of the Editor Board of Journal of Real-Time Image Processing (JRTIP).
- O. Sentieys is member of the editorial board of Journal of Low Power Electronics.
- M. Traiola and A. Kritikakou were guest co-editors of the Special Issue 'Dependability of Emerging Computing Paradigms and Technologies in IoT-Oriented Circuits, Architectures and Algorithms' in the MDPI Electronics journal in 2022.

9.1.4 Invited talks

- O. Sentieys gave an invited keynote at the Inria Scientific Days on ”Opportunities for computer architecture research with open-source hardware: The case for RISC-V“.
- D.Chillet gave a talk at RITS'22 (Recherche en Imagerie et Technologies pour la Santé), ”Carrière des Enseignants-Chercheurs”.
- A. Kritikakou gave an invited talk on ”Energy-Quality-Time Optimized Mapping for Imprecise Computation Tasks on Multicores” at a common thematic day on optimization methods organized by GDR RO and SOC2.
- M. Traiola gave an invited talk on “Approximate Computing in Hardware: Challenges and Opportunities for Test and Reliability” at the Winter French school on Design Technologies for Heterogeneous Embedded Systems (FETCH) 2022.

9.1.5 Leadership within the scientific community

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

9.1.6 Scientific expertise

- O. Sentieys was a member of the ANR Scientific Evaluation Committee CE25 "Software science and engineering - Multi-purpose communication networks, high-performance infrastructure".
- D. Chillet is a member of the HCERES Evaluation Committee for Master Programs of Université Nice Côte d'Azur.
- A. Kritikakou is a member of the Examination Committee of Industrial Engineering Sciences and Computer Engineering (SII) Aggrégation

9.1.7 Research administration

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

9.2 Teaching - Supervision

9.2.1 Teaching Responsibilities

- E. Casseau is in charge of the Department of “Digital Systems” at ENSSAT Engineering Graduate School.
- D. Chillet is associate director of studies at ENSSAT Engineering Graduate School.
- D. Chillet is the responsible of the ”Embedded Systems” major of the SISEA Master by Research.

- C. Killian is the responsible of the second year of the "Instrumentation" BUT at IUT, Lannion.
- S. Rokicki is the responsible of the second year in the computer science department of ENS Rennes

9.2.2 Teaching

- E. Casseau: programmable logic, 20h, 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, Master by Research (SISEA) and ENSSAT (M2)
- S. Derrien, optimizing and parallelising compilers, 14h, Master of Computer Science, ISTIC(M2)
- S. Derrien, advanced processor architectures, 8h, Master of Computer Science, ISTIC(M2)
- S. Derrien, high level synthesis, 20h, Master of Computer Science, ISTIC(M2)
- S. Derrien: introduction to operating systems, 8h, ISTIC (M1)
- S. Derrien, principles of digital design, 20h, Bachelor of EE/CS, ISTIC(L2)
- S. Derrien, computer architecture, 48h, Bachelor of Computer Science, ISTIC(L3)
- S.I. Filip, Operating Systems, 24h, Master of Mechatronics, ENS RENNES (M2)
- F. Charot: computer architecture, 48h, ESIR (L3)
- F. Charot: software hardware interfaces, 44h, ISTIC (L3)
- F. Charot: Compilation and code optimization architecture, 18h, ENSSAT (M2)
- D. Chillet: embedded processor architecture, 20h, ENSSAT (M1)
- D. Chillet: multimedia processor architectures, 24h, ENSSAT (M2)
- D. Chillet: advanced processor architectures, 20h, ENSSAT (M2)
- D. Chillet: micro-controller, 32h, ENSSAT (L3)
- D. Chillet: low-power digital CMOS circuits, 4h, UBO (M2)
- C. Killian: digital electronics, 75h, IUT Lannion (L1)
- C. Killian: automated measurements, 52.5h, IUT Lannion (L2)
- C. Killian: computer architecture, 6h, IUT Lannion (L3)
- C. Killian: embedded systems, 22.5h, IUT Lannion (L2)
- C. Killian: microcontrollers, 44h, IUT Lannion (L2)
- C. Killian: data acquisition and signal processing, 21h, UFAZ (L3)
- A. Kritikakou: principles of computer design, 32h, ISTIC (L3)
- A. Kritikakou: software hardware interfaces, 12h, ISTIC (L3)
- A. Kritikakou: C and unix programming languages, 76h, ISTIC (L3)
- A. Kritikakou: operating systems, 48h, ISTIC (L3)
- O. Sentieys: VLSI integrated circuit design, 24h, ENSSAT (M1)
- O. Sentieys: VHDL and logic synthesis, 18h, ENSSAT (M1)
- S. Rokicki: C Programming, 24h, ENS Rennes

9.2.3 PhD Supervision

- PhD: Davide Pala, Microarchitectures for Robust and Efficent Incremental Backup in Intermittently Powered Systems, Nov. 2022, O. Sentieys, I. Miro-Panades (CEA LETI).
- PhD: Minyu Cui, Energy-Quality-Time Fault Tolerant Task Mapping on Multicore Architectures, Jun. 2022, E. Casseau, A. Kritikakou.
- PhD: Jaechul Lee, Approximate communication techniques exploration for efficient nano-photonics interconnects, Dec. 2022, D. Chillet, C. Killian.
- PhD: Thibaut Marty, Timing speculation for hardware accelerators, March 2022, S. Derrien.
- PhD in progress: Thibault Allenet, Low-Cost Neural Network Algorithms and Implementations for Temporal Sequence Processing, March 2019, O. Sentieys, O. Bichler (CEA LIST).
- PhD in progress: 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.
- PhD in progress: Benoit Coqueret, Physical Security Attacks Against Artificial Intelligence Based Algorithms, Nov. 2022, O. Sentieys, M. Carbone (Thales), G. Zaid (Thales).
- PhD in progress: Minh Thanh Cong, Hardware Accelerated Simulation of Heterogeneous Multicore Platforms, May 2017, F. Charot, S. Derrien.
- PhD in progress: Léo De La Fuente, In-Memory Computing for Ultra Low Power Architectures, Nov. 2021, O. Sentieys, J.-F. Christmann (CEA LETI).
- 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: Van-Phu Ha, Application-Level Tuning of Accuracy, Nov. 2017, O. Sentieys.
- PhD in progress: Ibrahim Krayem, Fault tolerant emerging on-chip interconnects for manycore architectures, Oct. 2020, C. Killian, D. Chillet.
- PhD in progress: Seungah Lee, Efficient Designs of On-Board Heterogeneous Embedded Systems for Space Applications, Nov. 2021, A. Kritikakou, E. Casseau, R. Salvador (Centrale-Supelec), O. Sentieys.
- PhD in progress: 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 Pradels, Constrained optimization of FPGA accelerators for embedded deep convolutional neural networks, 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 Scientific production

10.1 Major publications

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

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

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

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

[5] S. Derrien, T. Marty, S. Rokicki and T. Yuki. 'Toward Speculative Loop Pipelining for High-Level Synthesis'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.11 (2020), pp. 4229–4239. DOI: 10.1109/TCAD.2020.3012866. URL: https://hal.archives-ouvertes.fr/hal-02949516.

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

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

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

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

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

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

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

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

[14] T. Marty, T. Yuki and S. Derrien. 'Safe Overclocking for CNN Accelerators through Algorithm-Level Error Detection'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.12 (Mar. 2020), pp. 4777–4790. DOI: 10.1109/TCAD.2020.2981056. URL: https://hal.inria.fr/hal-03094811.

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

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

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

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

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

10.2 Publications of the year

International journals

[20] M. Barbareschi, S. Barone, A. Bosio, J. Han and M. Traiola. 'A Genetic-algorithm-based Approach to the Design of DCT Hardware Accelerators'. In: *ACM Journal on Emerging Technologies in Computing Systems* 18.3 (31st July 2022), pp. 1–25. DOI: 10.1145/3501772. URL: https://hal.inria.fr/hal-03553505.

[21] M. Cui, A. Kritikakou, L. Mo and E. Casseau. 'Energy-Efficient Partial-Duplication Task Mapping under multiple DVFS schemes'. In: *International Journal of Parallel Programming* 50.2 (Apr. 2022), pp. 267–294. DOI: 10.1007/s10766-022-00724-7. URL: https://hal.inria.fr/hal-03907885.

[22] 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.

[23] F. Fernandes dos Santos, A. Kritikakou, J. Esteban Rodriguez Condia, J. David Guerrero Balaguera, M. Sonza Reorda, O. Sentieys and P. Rech. 'Characterizing a Neutron-Induced Fault Model for Deep Neural Networks'. In: *IEEE Transactions on Nuclear Science* (22nd Nov. 2022). URL: https://hal.inria.fr/hal-03865253.

[24] 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. 1–1. DOI: 10.1109/TCAD.2022.3201494. URL: https://hal.inria.fr/hal-03930715.

[25] J.-M. Gorius, S. Rokicki and S. Derrien. 'SpecHLS: Speculative Accelerator Design using High-Level Synthesis'. In: *IEEE Micro* (2022), pp. 1–10. DOI: 10.1109/mm.2022.3188136. URL: https://hal.inria.fr/hal-03714101.

[26] M. Kumm, A. Volkova and S.-I. Filip. 'Design of Optimal Multiplierless FIR Filters with Minimal Number of Adders'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (May 2022). DOI: 10.1109/TCAD.2022.3179221. URL: https://hal.science/hal-02392522.

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

[28] X. Li, L. Mo, A. Kritikakou and O. Sentieys. 'Approximation-aware Task Deployment on Heterogeneous Multi-core Platforms with DVFS'. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (15th Nov. 2022), pp. 1–15. DOI: 10.1109/TCAD.2022.3222293. URL: https://hal.science/hal-03854671.

[29] R. Psiakis, A. Kritikakou and O. Sentieys. 'Dynamic fault-tolerant VLIW processor with heterogeneous Function Units'. In: *Microprocessors and Microsystems: Embedded Hardware Design* 93 (25th May 2022), p. 104564. DOI: 10.1016/j.micpro.2022.104564. URL: https://hal.inria.fr/hal-03885490.

[30] J. Rodriguez Condia, P. Rech, F. Fernandes dos Santos, L. Carro and M. Sonza Reorda. 'An Effective Method to Identify Microarchitectural Vulnerabilities in GPUs'. In: *IEEE Transactions on Device and Materials Reliability* (30th Mar. 2022), pp. 1–14. DOI: 10.1109/TDMR.2022.3166260. URL: https://hal.inria.fr/hal-03669439.

[31] A. Ruospo, E. Sanchez, L. Matana Luza, L. Dilillo, M. Traiola and A. Bosio. 'A Survey on Deep Learning Resilience Assessment Methodologies'. In: *Computer* (2022). URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03834128.

International peer-reviewed conferences

[32] T. Allenet, D. Briand, O. Bichler and O. Sentieys. 'Disentangled Loss for Low-Bit Quantization-Aware Training'. In: CVPR 2022 - IEEE / CVF Computer Vision and Pattern Recognition Conference. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. New Orleans, United States, 2022, pp. 2788–2792. DOI: 10.1109/CVPRW56347.2022.00315. URL: https://hal-cea.archives-ouvertes.fr/cea-03776535.

[33] M. Barbareschi, A. Bosio, I. O'Connor, P. Fiser and M. Traiola. 'A Design Space Exploration Framework for Memristor-Based Crossbar Architecture'. In: DDECS 2022 - 25th International Symposium on Design and Diagnostics of Electronic Circuits and Systems. Prague, Czech Republic: IEEE, 6th Apr. 2022, pp. 38–43. DOI: 10.1109/DDECS54261.2022.9770145. URL: https://hal.inria.fr/hal-03888009.

[34] N. Bellec, G. Hiet, S. Rokicki, F. Tronel and I. Puaut. 'RT-DFI: Optimizing Data-Flow Integrity for Real-Time Systems'. In: ECRTS 2022 - 34th Euromicro Conference on Real-Time Systems. 34. Modène, Italy, 28th June 2022, pp. 1–24. DOI: 10.4230/LIPIcs.ECRTS.2022.18. URL: https://hal.inria.fr/hal-03641576.

[35] A. Bosio, M. Bragaglio, S. Germiniani, S. Mori, G. Pravadelli and M. Traiola. 'Assertion-aware approximate computing design exploration on behavioral models'. In: LATS 2022 - IEEE 23rd Latin American Test Symposium. Montevideo, Uruguay: IEEE, 5th Sept. 2022, pp. 1–6. DOI: 10.1109/LATS57337.2022.9936945. URL: https://hal.inria.fr/hal-03887690.

[36] 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://hal.inria.fr/hal-03887685.

[37] M. Cui, A. Kritikakou, L. Mo and E. Casseau. 'Near-optimal Energy-Efficient Partial-Duplication Mapping of Real-Time Parallel Applications'. In: AEiC 2022 - 26th Ada-Europe International Conference on Reliable Software Technologies. Ghent, Belgium, 14th June 2022, pp. 1–26. URL: https://hal.inria.fr/hal-03907727.

[38] N. Deligiannis, R. Cantoro, M. S. Reorda, M. Traiola and E. Valea. 'Improving the Fault Resilience of Neural Network Applications Through Security Mechanisms'. In: DSN-S 2022 - 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume. Baltimore, United States: IEEE, 27th June 2022, pp. 23–24. DOI: 10.1109/DSN-S54099.2022.00017. URL: https://hal.inria.fr/hal-03887704.

[39] F. Fernandes dos Santos, A. Kritikakou and O. Sentieys. 'Experimental evaluation of neutron-induced errors on a multicore RISC-V platform'. In: IOLTS 2022 - 28th IEEE International Symposium on OnLine Testing and Robust System Design. Torino, Italy: IEEE, 12th Sept. 2022, pp. 1–7. URL: https://hal.inria.fr/hal-03697265.

[40] F. Fernandes dos Santos and P. Rech. 'Performance-Reliability Trade-Off in Graphics Processing Units'. In: RADiation Effects on Components and Systems (RADECS). Venice, Italy, 3rd Oct. 2022. URL: https://hal.inria.fr/hal-03680872.

[41] J.-M. Gorius, S. Rokicki and S. Derrien. 'Design Exploration of RISC-V Soft-Cores through Speculative High-Level Synthesis'. In: *2022 International Conference on Field-Programmable Technology (ICFPT)*. FPT 2022 - International Conference on Field Programmable Technology. Honk Kong / Hybrid, Hong Kong SAR China: IEEE, 2022, pp. 1–6. DOI: 10.1109/ICFPT56656.2022.9974478. URL: https://hal.inria.fr/hal-03828841.

[42] 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, 17th Apr. 2023. URL: https://hal.inria.fr/hal-03885240.

[43] I. Krayem, R. Mercier, C. Killian, A. Kritikakou and D. Chillet. 'Data and Fault Aware Routing Algorithm for NoC Based Approximate Computing'. In: NANOARCH 2022 - 17th ACM International Symposium on Nanoscale Architectures. Virtual, France: ACM, 7th Dec. 2022, pp. 1–6. URL: https://hal.science/hal-03920728.

[44] A. Kritikakou, P. Nikolaou, I. Rodriguez-Ferrandez, J. Paturel, L. Kosmidis, M. K. Michael, O. Sentieys and D. Steenari. 'Functional and Timing Implications of Transient Faults in Critical Systems'. In: IOLTS 2022 – 28th IEEE International Symposium on OnLine Testing and Robust System Design. Torino, Italy: IEEE, 12th Sept. 2022, pp. 1–10. URL: https://hal.science/hal-03923506.

[45] A. Kritikakou, O. Sentieys, G. Hubert, Y. Helen, J.-F. Coulon and P. Deroux-Dauphin. 'FLODAM: Cross-Layer Reliability Analysis Flow for Complex Hardware Designs'. In: DATE 2022 - 25th IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium: IEEE, 14th Mar. 2022, pp. 1–6. URL: https://hal.science/hal-03485386.

[46] R. Mercier, C. Killian, A. Kritikakou, Y. Helen and D. Chillet. 'Tolerating Errors in NoC: A Lightweight Region-Based Fault-Mitigation Method'. In: SELSE 2022 - IEEE Workshop on Silicon Errors in Logic – System Effects. [Virtual], France: IEEE, 19th May 2022, pp. 1–7. URL: https://hal.inria.fr/hal-03926148.

[47] L. Mo, Q. Zhou, A. Kritikakou and J. Liu. 'Energy Efficient, Real-time and Reliable Task Deployment on NoC-based Multicores with DVFS'. In: DATE 2022 - IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium: IEEE, 14th Mar. 2022, pp. 1–6. URL: https://hal.science/hal-03500332.

[48] V. L. Nguyen Huu, J. Lallet, E. Casseau and L. d'Orazio. 'Cache management in MASCARA-FPGA: from coalescing heuristic to replacement policy'. In: DaMoN 2022 - 18th International Workshop on Data Management on New Hardware. Philadelphia, United States: ACM, 13th June 2022, pp. 1–5. DOI: 10.1145/3533737.3535096. URL: https://hal.inria.fr/hal-03907912.

[49] A. Piri, S. Saeedi, M. Barbareschi, B. Deveautour, S. D. Carlo, I. O'Connor, A. Savino, M. Traiola and A. Bosio. 'Input-Aware Approximate Computing'. In: AQTR 2022 - IEEE International Conference on Automation, Quality and Testing, Robotics. Cluj-Napoca, Romania: IEEE, 19th May 2022, pp. 1–6. DOI: 10.1109/AQTR55203.2022.9801944. URL: https://hal.inria.fr/hal-03887997.

[50] M. Rajan, A. Das and J. Jose. 'LOKI: A Hardware Trojan Affecting Multiple Components of an SoC'. In: 2022 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). Paphos, Cyprus, 4th July 2022. DOI: 10.1109/ISVLSI54635.2022.00043. URL: https://hal.science/hal-03676968.

[51] A. Ruospo, G. Gavarini, I. Bragaglia, M. Traiola, A. Bosio and E. Sanchez. 'Selective Hardening of Critical Neurons in Deep Neural Networks'. In: DDECS 2022 - 25th International Symposium on Design and Diagnostics of Electronic Circuits and Systems. Prague, Czech Republic: IEEE, 6th Apr. 2022, pp. 136–141. DOI: 10.1109/DDECS54261.2022.9770168. URL: https://hal.inria.fr/hal-03888005.

[52] F. F. dos Santos, P. Rech, A. Kritikakou and O. Sentieys. 'Evaluating the Impact of Mixed-Precision on Fault Propagation for Deep Neural Networks on GPUs'. In: ISVLSI 2022 - IEEE Computer Society Annual Symposium on VLSI. Nicosia, Italy: IEEE, 4th July 2022, pp. 327–327. DOI: 10.1109/ISVLSI54635.2022.00071. URL: https://hal.inria.fr/hal-03903347.

[53] M. Tatsumi, S.-I. Filip, C. White, O. Sentieys and G. Lemieux. 'Mixing Low-Precision Formats in Multiply-Accumulate Units for DNN Training'. In: FPT 2022 - IEEE International Conference on Field Programmable Technology. Hong Kong, Hong Kong SAR China: IEEE, 5th Dec. 2022, pp. 1–9. URL: https://hal.inria.fr/hal-03885471.

[54] M. Traiola, A. Kritikakou and O. Sentieys. 'harDNNing: a machine-learning-based framework for fault tolerance assessment and protection of Deep Neural Networks'. In: DATE 2023 – 26th IEEE/ACM Design, Automation and Test in Europe. Antwerp, Belgium, 17th Apr. 2023. URL: https://hal.inria.fr/hal-03887681.

[55] P. Zolfaghari, J. Ortiz, C. Killian and S. Le Beux. 'Non-Volatile Phase Change Material based Nanophotonic Interconnect'. In: DATE 2022 - IEEE/ACM Design, Automation and Test in Europe. IEEE/ACM Design, Automation and Test in Europe (DATE) 2022. Antwerp, Belgium: IEEE, 1st Dec. 2021, pp. 1–6. URL: https://hal.science/hal-03512251.

National peer-reviewed Conferences

[56] R. Mercier, C. Killian, A. Kritikakou, Y. Helen and D. Chillet. 'Atténuation des Défauts dans les Réseaux sur Puce avec une Approche de Brassage de Bits Basée sur des Régions'. In: GRETSI 2022 - XXVIIIème Colloque Francophone de Traitement du Signal et des Images. Nancy, France, 6th Sept. 2022, pp. 1–4. URL: https://hal.inria.fr/hal-03926136.

Conferences without proceedings

[57] F. Fernandes dos Santos, A. Kritikakou, O. Sentieys and P. Rech. 'Characterizing Deep Neural Networks Neutrons-Induced Error Model'. In: NSREC 2022 - IEEE Nuclear & Space Radiation Effects Conference. Provo, United States: IEEE, 18th July 2022, pp. 1–5. URL: https://hal.inria.fr/hal-03652138.

Scientific books

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

Scientific book chapters

[59] A. Bosio, S. Di Carlo, P. Girard, A. Ruospo, E. Sanchez, A. Savino, L. Sekanina, M. Traiola, Z. Vašíček and A. Virazel. 'Design, Verification, Test, and In-Field Implications of Approximate Digital Integrated Circuits'. In: *Approximate Computing Techniques*. Springer International Publishing, 3rd Jan. 2022, pp. 349–385. DOI: 10.1007/978-3-030-94705-7_12. URL: https://hal.inria.fr/hal-03888027.

[60] E. Dupuis, S.-I. Filip, O. Sentieys, D. Novo, I. O'Connor and A. Bosio. 'Approximations in Deep Learning'. In: *Approximate Computing Techniques - From Component- to Application-Level*. 2022, pp. 467–512. DOI: 10.1007/978-3-030-94705-7_15. URL: https://hal.science/hal-03494874.

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

[62] M. Traiola, B. Deveautour, A. Bosio, P. Girard and A. Virazel. 'Test and Reliability of Approximate Hardware'. In: *Approximate Computing*. Springer International Publishing, 18th Mar. 2022, pp. 233–266. DOI: 10.1007/978-3-030-98347-5_10. URL: https://hal.inria.fr/hal-03888016.

Doctoral dissertations and habilitation theses

[63] M. Cui. 'Energy-Quality-Time Fault Tolerant Task Mapping on Multicore Architectures'. École normale supérieure de Rennes, 24th June 2022. URL: https://theses.hal.science/tel-03765873.

[64] C. Killian. 'Energy efficiency, fault tolerance, and emerging on-chip interconnects for manycore architectures'. Université de Rennes 1 (UR1), 13th June 2022. URL: https://hal.science/tel-03949792.

[65] A. Kritikakou. 'Real-time and reliable design for safety-critical embedded systems'. Rennes 1, 28th Nov. 2022. URL: https://hal.science/tel-03965028.

[66] J. Lee. 'Approximate communication techniques exploration for efficient nano-photonics interconnects'. Université Rennes 1, 8th Dec. 2022. URL: https://theses.hal.science/tel-03958287.

[67] T. Marty. 'Temporal Speculation for hardware accelerators'. Université de Rennes 1, 24th Mar. 2022. URL: https://hal.inria.fr/tel-03925783.

[68] D. Pala. 'Microarchitectures for Robust and Efficient Incremental Backup in Intermittently-Powered Systems'. Université de Rennes 1, 10th Nov. 2022. URL: https://hal.inria.fr/tel-03885206.

Reports & preprints

[69] N. Cavagnero, F. Fernandes dos Santos, M. Ciccone, G. Averta, T. Tommasi and P. Rech. *Fault-Aware Design and Training to Enhance DNNs Reliability with Zero-Overhead*. 1st June 2022. URL: https://hal.inria.fr/hal-03684224.

Other scientific publications

[70] F. Fernandes, A. Kritikakou and O. Sentieys. 'Experimental evaluation of neutron-induced errors on a RISC-V processor'. In: RISC-V Week 2022. Paris, France, 3rd May 2022. URL: https://hal.inria.fr/hal-03903370.

[71] S. Rokicki, J. Paturel and O. Sentieys. 'Comet: a RISC-V Core Synthesized from C++ Specifications'. In: Spring 2022 RISC-V Week. Paris, France, 3rd May 2022. URL: https://hal.inria.fr/hal-03885663.

10.3 Cited publications

[72] A. P. 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

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

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

[84] 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.

[85] 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.