RESEARCH CENTRE

Inria Centre at Rennes University

IN PARTNERSHIP WITH:

École normale supérieure de Rennes, Université de Rennes

2024
ACTIVITY REPORT

Project-Team TARAN

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

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

DOMAIN

Algorithmics, Programming, Software and Architecture

THEME

Architecture, Languages and Compilation

Contents

Project-Team TARAN

Creation of the Project-Team: 2021 May 01

Keywords

Computer sciences and digital sciences

A1.1. – Architectures

A1.1.1. – Multicore, Manycore

A1.1.2. – Hardware accelerators (GPGPU, FPGA, etc.)

A1.1.8. – Security of architectures

A1.1.9. – Fault tolerant systems

A1.1.10. – Reconfigurable architectures

A1.1.12. – Non-conventional architectures

A1.2.5. – Internet of things

A1.2.6. – Sensor networks

A2.2. – Compilation

A2.2.4. – Parallel architectures

A2.2.6. – GPGPU, FPGA...

A2.2.7. – Adaptive compilation

A2.2.8. – Code generation

A2.3.1. – Embedded systems

A2.3.3. – Real-time systems

A4.4. – Security of equipment and software

A8.10. – Computer arithmetic

A9.9. – Distributed AI, Multi-agent

Other research topics and application domains

B4.5. – Energy consumption

B4.5.1. – Green computing

B4.5.2. – Embedded sensors consumption

B6.4. – Internet of things

B6.6. – Embedded systems

1 Team members, visitors, external collaborators

Research Scientists

- Olivier Sentieys [Team leader, INRIA, Senior Researcher]
- François Charot [INRIA, Researcher, until Jun 2024]
- Fernando Fernandes Dos Santos [INRIA, ISFP]
- Silviu-Ioan Filip [INRIA, Researcher]
- Marcello Traiola [INRIA, Researcher]

Faculty Members

- Emmanuel Casseau [UNIV RENNES, Professor, until Sep 2024, HDR]
- Daniel Chillet [UNIV RENNES, Professor, HDR]
- Steven Derrien [UNIV RENNES, Professor, until Aug 2024, HDR]
- Angeliki Kritikakou [UNIV RENNES, Associate Professor, HDR]
- Bertrand Le Gal [UNIV RENNES, Associate Professor, HDR]
- Patrice Quinton [ENS RENNES, Emeritus, HDR]
- Simon Rokicki [ENS RENNES, Associate Professor]

Post-Doctoral Fellows

- Rafael Billig Tonetto [UNIV RENNES, Post-Doctoral Fellow, from May 2024]
- El-Mehdi El Arar [UNIV RENNES, Post-Doctoral Fellow, from Feb 2024]
- Remi Garcia [UNIV RENNES, Post-Doctoral Fellow]

PhD Students

- Oussama Ait Sidi Ali [SAFRAN, CIFRE]
- Hamza Amara [UNIV RENNES]
- Herinomena Andrianatrehina [INRIA]
- Gaetan Barret [ORANGE, CIFRE]
- Sami Ben Ali [INRIA]
- Benoit Coqueret [THALES, CIFRE]
- Leo De La Fuente [CEA, until Nov 2024]
- Sohaib Errabii [INRIA, from May 2024]
- Romain Facq [INRIA, from Oct 2024]
- Corentin Ferry [UNIV RENNE, until May 2024]
- Jean-Michel Gorius [UNIV RENNES, until Nov 2024]
- Wilfread Guilleme [INRIA]

- Ibrahim Krayem [UNIV RENNES, until Feb 2024]
- Seungah Lee [UNIV RENNES, until Oct 2024]
- Dylan Leothaud [UNIV RENNES]
- Guillaume Lomet [INRIA]
- Amélie Marotta [INRIA]
- Saptarshi Nag [INRIA, until Jan 2024]
- Louis Narmour [UNIV RENNES, until Sep 2024]
- Pegdwende Romaric Nikiema [UNIV RENNES]
- Leo Pajot [KEYSOM SAS, CIFRE]
- Leo Pradels [INRIA, until Mar 2024]
- Lucas Roquet [UNIV RENNES, from Oct 2024]
- Baptiste Rossigneux [CEA]
- Louis Savary [INRIA]
- Nesrine Sfar [UNIV RENNES, from Dec 2024]
- Anis Yagoub [KEYSOM SAS, from Aug 2024]

Technical Staff

- Ibrahim Krayem [UNIV RENNES, Engineer, from Apr 2024 until May 2024]
- Joseph Paturel [INRIA, Engineer]
- Dikshanya Lashmi Ramaswamy [INRIA, Engineer, until Feb 2024]
- Etienne Tehrani [INRIA, Engineer]

Interns and Apprentices

- El Mehdi Bel Haddad [INRIA, Intern, from Sep 2024]
- Estephe Beyriere [UNIV RENNES, Intern, from Apr 2024 until Sep 2024]
- Nour Chiboub [INRIA, Intern, from Jun 2024 until Nov 2024]
- Lucas Roquet [UNIV RENNES, Intern, from Mar 2024 until Aug 2024]
- Nesrine Sfar [INRIA, Intern, from Mar 2024 until Aug 2024]

Administrative Assistants

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

Visiting Scientist

- Alessio Colucci [UNIV VIENNE, from Apr 2024 until May 2024]

External Collaborator

- Guillaume Didier [DGA]

2 Overall objectives

Energy efficiency has now become one of the main requirements for virtually all computing platforms [84]. 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 [83, 82], 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 [78]. The main driver today is energy efficiency—small embedded devices need specialized hardware to operate under power/energy constraints. In the next ten years, we expect specializations to become even more common to meet increasing demands for performance. In particular, high-throughput workloads traditionally run on servers (e.g., computational science and machine learning) will offload (parts of) their computations to accelerators. We are already seeing some instances of such specialization, most notably accelerators for neural networks that use clusters of nodes equipped with FPGAs and/or ASICs.

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

2.2 Design Stack for Custom Hardware

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

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

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

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

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

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

2.3 Objectives of TARAN: Facilitating Cross-Layer Optimization

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

3 Research program

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

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

The common keyword across all directions is **domain-specific**. Specialization is necessary for addressing various challenges including productivity, efficiency, reliability, and scalability in the next generation of computing platforms. Our main objective is defined by the need to jointly work on multiple layers of the design stack to be truly domain-specific. Another common challenge for the entire team is **design**

space exploration, which has been and will continue to be an essential process for HW design. We can only expect the design space to keep expanding, and we must persist on developing techniques to efficiently navigate through the design space.

3.1 Accelerators

Key Investigators: E. Casseau, F. Charot, D. Chillet, S. Derrien, A. Kritikakou, B. Le Gal, P. Quinton, S. Rokicki, O. Sentieys.

Accelerators are custom hardware that primarily aim to provide high-throughput, energy-efficient, computing platforms. Custom hardware can give much better performance compared to more general architectures simply because they are specialized, at the price of being much harder to "program." Accelerator designers need to explore a massive design space, which includes many hardware parameters that a software programmer has no control over, to find a suitable design for the application at hand.

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

Positioning: Hardware acceleration has already proved its efficiency in many datacenter, cloud-computing or embedded high-performance computing (HPC) applications: machine learning, web search, data mining, database access, information security, cryptography, financial, image/signal/video processing, etc. For example, the work at Microsoft in accelerating the Bing web search engine with large-scale reconfigurable fabrics has shown to improve the ranking throughput of each server by 95% [88], and the increasing need for acceleration of deep learning workloads [91].

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 [79, 86]. On this topic, we have been pursuing our efforts to propose tools whose design principles are based on a tight coupling between the compiler and the target hardware architectures.

3.2 Accurate Computing

Key Investigators: S. Filip, S. Derrien, O. Sentieys, M. Traiola.

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

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

Positioning: Traditionally, fixed-point (FxP) arithmetic is used to relax accuracy, providing important benefits in terms of delay, power and area [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 [87]. 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 [80].

3.3 Resilient Computing

Key Investigators: E. Casseau, D. Chillet, F. Fernandes dos Santos, A. Kritikakou, O. Sentieys, M. Traiola. With advanced technology nodes and the emergence of new devices pressured by the end of Moore's law, manufacturing problems and process variations strongly influence electrical parameters of circuits and architectures [85], leading to dramatically reduced yield rates [89]. Transient errors caused by particles or radiations will also more and more often occur during execution [92, 90], 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, O. Sentieys, M. Traiola.
Domain specific accelerators have more exploratory freedom to take advantage of non-conventional technologies that are too specialized for general purpose use. Examples of such technologies include optical interconnects for Network-on-Chip (NoC) and Non-Volatile Memory (NVM) for low-power sensor nodes. The objective of this research direction is to explore the use of such technologies, and find appropriate application domains. The primary cross-layer interaction is expected from Hardware Design to accommodate non-conventional Technologies. However, this research direction may also involve Runtime and Compilers.

4 Application domains

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

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

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

5 Highlights of the year

5.1 Awards

- A. Kritikakou received the outstanding TPC member award of Design Automation Conference (DAC) 2024
- A. Kritikakou and M. Traiola received recognition for “significant contributions over time to the IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS)"

6 New software, platforms, open data

6.1 New software

6.1.1 Gecos

Name: Generic Compiler Suite

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

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

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

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

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

Publication: hal-03714101

Contact: Steven Derrien

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

Partner: Université de Rennes 1

6.1.2 SmartSense

Name: Sensor-Aided Non-Intrusive Load Monitoring

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

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

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

Contact: Olivier Sentieys

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

6.1.3 TypEx

Name: Type Exploration Tool

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

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

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

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

Contact: Olivier Sentieys

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

6.2 New platforms

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

Participants: Silviu-Ioan Filip.

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

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

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

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

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

Participants: Silviu-Ioan Filip.

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

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

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

6.2.3 Hybrid-DBT

Participants: Simon Rokicki, Louis Savary, Steven Derrien.

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

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

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

6.2.4 Comet

Participants: Simon Rokicki, Olivier Sentieys, Joseph Paturel.

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

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

Comet is still in active development. Our roadmap includes 64-bit version, Linux compatible, support for vector ISA extension, out-of-order superscalar microarchitecture. Comet is used in many ongoing research projects (PEPR Arsene, Cyberpros, Foch, CocoRISCo Inria Challenge, EuroHPC DARE (Digital Autonomy with RISC-V in Europe)) and is used as support to several PhD theses.

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

6.2.5 MMAlpha

Participants: Patrice Quinton.

KEYWORDS: High-level synthesis, polyhedral model

SCIENTIFIC DESCRIPTION: `MMAlpha` is a software for the high-level synthesis of parallel architectures from high-level specifications written using the Alpha language. Developed over years, it is currently able to generate automatically synthesizable Vhdl programs for various examples. MMAlpha was recently used to generate Vhdl code for the simulation of electrical circuits.

- Contact: Patrice Quinton
- Partners: Univ Rennes
- URL: none

7 New results

7.1 High-Level Synthesis of Speculative Hardware Accelerators

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

High-Level Synthesis (HLS) is a powerful method for generating hardware accelerators from C/C++ code, making it easier to develop complex digital circuits. However, current HLS tools have limitations when it comes to efficiently scheduling loops, especially with complex control-flow and memory dependencies. These limitations prevent HLS from fully exploiting the parallelism in modern applications, which impacts the efficiency of the generated accelerators.

The technique we have developed in this project is Speculative Loop Pipelining [5], which allows us to generate speculative accelerators that are more complex than manually designed ones, and even processors. In 2024, our work continued to push the boundaries of speculative loop pipelining, focusing on improving the SpecHLS toolchain to address these challenges. Specifically, we contributed to Memory Speculation Management and Design Space Exploration.

In [44], we present an approach to manage memory speculation. This involves handling arrays during speculative execution, including rolling back changes when speculation fails. We extend the speculative loop pipelining framework to manage inter-iteration memory dependencies, allowing more aggressive scheduling in the HLS process while ensuring correctness. The paper also describes backend transformations to generate an efficient and correctly sized hardware structure, similar to a load-store queue used in modern out-of-order processors.

In [50] and [61], we address the challenge of efficiently exploring the design space for speculative loop pipelining. The key issue is deciding where to apply speculation, as the number of possible combinations can grow exponentially with the number of opportunities available. For example, the Instruction Set Simulator (ISS) of a RISC-V processor presents around twenty different speculative opportunities, leading to a very complex design space. To address this, we propose a method that efficiently narrows down the design space, focusing on configurations that achieve an initiation interval of one. Our approach uses static profiling to estimate the probability of mispeculation and topological analysis to determine if a given speculation configuration can achieve an initiation interval of one when combined with other speculations. This helps ensure that the SpecHLS tool can effectively handle real-world applications. The project has also been presented at the European RISC-V Summit in Munich [75].

7.2 High-Level Synthesis-Based On-board Payload Data Processing

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

The complexity of on-board data processing platforms has increased with the development of heterogeneous System-on-Chips (SoCs) in the aerospace engineering domain. As space scientists need to implement computation-intensive algorithms on on-board platforms, it is a complex task for non-experts in embedded development to select proper hardware such as a CPU or an FPGA in a given processing platform. In [49], we propose PADIA as a Design Space Exploration (DSE) tool to provide estimations of ideal hardware models to implement user's algorithms on heterogeneous embedded systems. PADIA uses the LLVM intermediate representation and LLVM's analysis and transform passes to perform DSE targeting different hardware configurations simultaneously. Especially, High-Level Synthesis-based optimizations are used for FPGA designs. The tool internally generates matrices, based on LLVM instructions and optimizations, regarding user's input algorithms and pre-defined algorithms specifically crafted for CPUs and FPGAs. Then, the tool analyzes the similarity of the generated matrices to find the best hardware target. From DSE results, PADIA provides not only an expected proper hardware candidate, but also estimations on efficient area and latency optimizations. This work is done in collaboration with CNES (France) who initiated PADIA before, but effective LLVM-based optimizations were missing because the tool applied only one LLVM transform pass per optimization.

7.3 Training Deep Neural Networks with Low-Precision Accelerators

Participants: Sami Ben Ali, El-Mehdi El Arar, Silviu Filip, Olivier Sentieys.

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

Recent work has aimed to mitigate this computational challenge by introducing 8-bit floating-point (FP8) formats for multiplication. However, accumulations are still done in either half (16-bit) or single (32-bit) precision arithmetic. In [37], we investigate lowering accumulator word length while maintaining the same model accuracy. We present a multiply-accumulate (MAC) unit with FP8 multiplier inputs and FP12 accumulations, which leverages an optimized stochastic rounding (SR) implementation to mitigate swamping errors that commonly arise during low precision accumulations. We empirically investigate the hardware implications and accuracy impact associated with varying the number of random bits used for rounding operations. We additionally attempt to reduce MAC area and power by proposing a new scheme to support SR in floating-point MAC and by removing support for subnormal values. Our optimized eager SR unit significantly reduces delay and area when compared to a classic lazy SR design. Moreover, when compared to MACs utilizing single-or half-precision adders, our design showcases notable savings in all metrics. Furthermore, our approach consistently maintains near baseline accuracy across a diverse range of computer vision tasks, making it a promising alternative for low-precision DNN training.

In a follow-up work [71], we look at the error analysis implications of using SR with a fixed number of random bits. We derive a probabilistic error model based on martingales and the Bienaymé-Chebyshev inequality that allows us to derive a theoretically sound heuristic for choosing the number of random bits used in long computation chains involving SR. This result is verified on several computational examples relevant to scientific computing, numerical optimization and large scale machine learning.

Several frameworks explore custom number formats with parameterizable precision through software emulation on CPUs or GPUs. However, they lack comprehensive support for different rounding modes and struggle to accurately evaluate the impact of custom precision for FPGA-based targets. In [38], we introduce MPTorch-FPGA, an extension of our MPTorch framework for performing custom, multi-precision inference and training computations in CPU, GPU, and FPGA environments in PyTorch. MPTorch-FPGA can generate multiple systolic arrays, each with independent sizes and custom arithmetic implementations that directly provide bit-level accuracy to accelerate GEMM calculations by offloading from the CPU or GPU. An offline matching algorithm selects one of several pre-generated (static) FPGA configurations using a custom performance model to estimate latency. A series of training benchmarks using diverse DNN models are explored, with a wide range of number format configurations and rounding modes. We report both accuracy and hardware performance metrics, verifying the precision of our performance model by comparing estimated and measured latencies across multiple benchmarks. These results highlight the flexibility and practical value of our framework.

Part of this work is conducted in collaboration with University of British Columbia, Vancouver, Canada and University of Leeds, United Kingdom.

7.4 Compression for DNN Inference

Participants: Cédric Gernigon, Léo Pradels, Baptiste Rossigneux, Silviu Filip, Olivier Sentieys, Daniel Chillet, Emmanuel Casseau.

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

In [28], we propose various algorithms for model compression by exploiting weight characteristics and conducts an in-depth study of their performance. The algorithms involve manipulating exponents and mantissa in the floating-point representations of weights. In addition, we also present a retraining method that uses the proposed algorithms to further reduce the size of pre-trained models. The results presented in this article are mainly on BFloat16 floating-point format. The proposed weight manipulation algorithms save at least 20% of memory on state-of-the-art image classification models compared to BFloat16 floating-point with very minor accuracy loss. This loss is bridged using the retraining method that saves at least 30% of memory, with potential memory savings of up to 43%. We compare the performance of the proposed methods against the state-of-the-art model compression techniques in terms of accuracy, memory savings, inference time, and energy. This work is conducted in collaboration with IIT Goa, India.

Model quantization is a common approach to deal with deployment constraints, but searching for optimized bit-widths can be challenging. In [42], we present Adaptive Bit-Width Quantization Aware Training (AdaQAT), a learning-based method that automatically optimizes weight and activation signal bit-widths during training for more efficient DNN inference. We use relaxed real-valued bit-widths that are updated using a gradient descent rule, but are otherwise discretized for all quantization operations. The result is a simple and flexible QAT approach for mixed-precision uniform quantization problems. Compared to other methods that are generally designed to be run on a pretrained network, AdaQAT works well in both training from scratch and fine-tuning scenarios. Initial results on the CIFAR-10 and ImageNet datasets using ResNet20 and ResNet18 models, respectively, indicate that our method is competitive with other state-of-the-art mixed-precision quantization approaches. In [66], we have further extended this method to handle finer levels of quantization granularity (*i.e.,* at the layer and sub-layer levels) using a block coordinate descent optimization approach, leading to equally impressive and state-of-the-art equivalent results, at a fraction of the time required by other methods in this space. This work is conducted in collaboration with CNES.

While convolutional neural networks (CNNs) have demonstrated exceptional performance in computer vision, optimizing FPGA-based CNN accelerators remains a challenge due to resource constraints.

This is especially true for inter-layer pipelining, which limits external memory access. Despite the benefits of sparsity, most existing sparse accelerators are sequential and memory-bound. In [54], we introduce an innovative inter-layer pipelined CNN architecture enriched with structured sparsity through pattern pruning. In our approach, pattern pruning serves as a fine-tuning step, effectively reducing FPGA resource consumption, including memory and logic. Experimental results indicate that our method leads to better latency than other inter-layer pipelining approaches, but also maintains competitive accuracy compared to state-of-the-art unstructured pruning methods. We demonstrate the versatility of our approach in image classification and super-resolution applications, achieving a consistent 30 frames per second across a wide range of image sizes on the Set5 dataset. This work was done in collaboration with Safran Electronics & Defense.

Additionally, as CNNs consist of successions of linear and nonlinear operations, we propose a new procedure to linearize CNNs. We leverage information from the inputs to each nonlinear functions to identify which nonlinearities are less critical for the network's performance. Our method is versatile, adaptable to any common nonlinearity and CNN architecture. While it gives a small drop in accuracy across a wide range of CNNs with respect to state-of-the-art methods, it by-passes the usual significant computational effort to determine removable nonlinearities, whether layer-wise or channel-wise. This work is done in collaboration with CEA LIST (France).

7.5 Design of Hardware Accelerators based on Approximate Computing (AxC)

Participants: Marcello Traiola.

Various Approximate Computing (AxC) techniques have been proposed so far in the literature at different abstraction levels, from hardware to software. These techniques have been successfully utilized and combined to realize approximate implementations of applications in various domains (e.g., data analytics, scientific computing, multimedia and signal processing, and machine learning). Usually, approximation methodologies focus on a single abstraction level, such as elementary components (e.g., arithmetic operations), and only then are combined at the application level. We designed and implemented a design framework to provide an application-driven approximation approach targeting different implementations (i.e., hardware and software). The approach automatically generates approximate variants of applications and performs a Design-Space Exploration (DSE) to find the available accuracy-efficiency trade-offs [64]. As the design exploration space quickly becomes the bottleneck for successfully deploying AxC, we investigated a methodology to identify resilient elements (e.g, hardware component, HDL statements) of the design to be approximated. By leveraging an assertion-based verification methodology in combination with fault injection, we can guide the AxC DSE of RTL descriptions [23]. Activities are starting on Approximate Fault-Tolerant hardware systems [57, 36], at the intersection between the Accurate and Resilient computing axes of the team.

7.6 Compiler Optimization Impact on GPU Error Rate

Participants: Fernando Fernandes dos Santos.

Graphics Processing Units (GPUs) compilers have adapted to support general-purpose programming across multiple micro-architectures. The NVIDIA CUDA Compiler (NVCC) incorporates multiple compilation flags and complex optimizations that impact not just performance but also GPU reliability. In [24, 26], we evaluated the effect of NVCC optimization flags on GPU error rates using two NVIDIA architectures (Kepler and Volta) and two compiler versions (10.2 and 11.3). By analyzing eight workloads across 144 compilation flag combinations, we found that optimizations can significantly influence error rates. Notably, experiments showed that the unoptimized GEMM had a lower error rate than its optimized counterparts. When the performance is evaluated together with the error rate, we show that the most optimized versions always produce a higher amount of correct data than the unoptimized code.

7.7 Reliability of ANN Hardware Accelerators

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

Dedicated hardware is essential for efficiently executing the resource-intensive modern artificial neural networks (ANNs). The increasing complexity of these ANNs has led to adopting sophisticated frameworks that generate optimized code for hardware accelerators such as GPUs and facilitate the creation of actual hardware accelerators on FPGAs using high-level synthesis (HLS) tools. These high abstractions simplify the software and hardware development process for programmers and designers, complicating accurate reliability assessments. Furthermore, the size of modern ANNs has surged at an unprecedented rate, necessitating ever-larger hardware accelerators. We evaluated the dependability of such large DNN accelerators [39, 47, 43]. In particular:

We evaluated the failure rate of ANN hardware accelerators generated by HLS tools under high-energy neutrons and explored the impact of HLS parameters on reliability. The results in [32] show that tweaking hardware parameters, such as the reuse of resources, can increase the error rate linearly. The generated ANN hardware accelerator with the best tradeoff of area and execution cycles can deliver 15× more correct executions than the least optimized one despite its increased error rate.

Using a neutron beam, we also assessed the failure rate and fault model of large Vision Transformers (ViTs). We identified the faults most likely propagating to the output and developed tailored procedures that are efficiently integrated into the ViT to locate and correct these faults. We proposed *MaxiMum corrupted values* (MaxiMals), an experimentally tuned low-cost mitigation solution to reduce the impact of transient faults on ViTs. The results in [55] demonstrated that MaxiMals can correct 90.7% of critical failures, with execution time overheads as low as 5.61%.

Although traditional reliability methods such as MaxiMals, hardware and software replications, and partial protection, have a high protection efficiency, they generally come with overhead costs. In [25, 48, 27] we proposed *Design ImprovEd HARDened neural Network* (DieHardNet), a specialized DNN designed with various hardening techniques to enhance robustness against transient faults. Those hardening methods were applied at the design and training time, incurring no overhead at the inference. We evaluated these strategies through extensive experiments on vision classification tasks. Results show that DieHardNet can reduce the critical error rate by up to 100 times compared to unprotected models.

Other contributions on the reliability of ANN hardware accelerators include [22].

7.8 Reliability Assessment of RISC-V Processors

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

The RISC-V Instruction Set Architecture (ISA) has gained popularity among systems designers thanks to its open-source nature. Its high flexibility has allowed it to be preferred in various domains and used to target multiple use cases, from embedded systems as co-processor to high-performance computers. Embedded systems, in general, and safety-critical ones, in particular, have strict requirements in terms of reliability and availability. The hardware is becoming less robust with the adoption of smaller technology nodes. The smaller transistor size, low operating voltage, and high switching frequency make transistors susceptible to Single-Event Upsets (SEU) faults, which can propagate to the application output and possibly cause catastrophic consequences.

RISC-V architectures can be customized to efficiently run Machine Learning (ML) algorithms in safety-critical domains. However, hardware faults can compromise system performance. Therefore, it is essential to characterize the vulnerabilities of ML applications on RISC-V processors and assess how errors impact the CNN misclassification rate. In [40], we evaluate the error rate induced by neutrons on CNN operations running on a RISC-V-based processor (GAP8) and analyze each operation's contribution to the overall error rate. Our findings show that memory errors primarily affect the system's error rate. We also present a case study illustrating how CNN microbenchmarks can estimate the error rate of an entire

CNN. By combining fault simulation data and beam experiments, our estimation closely matches the results from beam experiments alone.

Furthermore, during the software design phase of the system, compilation optimizations can be made to improve the performance. Compilers have various flags that modify the source code to produce the binary. Although these flags can be crucial in assuring good performance, they can significantly impact the resilience to SEU. In [52], we provide comprehensive insights into the impact of compiler optimizations on the reliability of safety-critical embedded systems. Specifically, a probabilistic fault injection campaign is conducted on various benchmarks running on a RISC-V core to evaluate the effect of several optimizations on reliability. The results are classified into functional and timing errors, offering a detailed understanding of the implications of these optimizations on reliability.

7.9 Reliability-Aware and Energy Efficient Task Mapping on NoC-Based MPSoCs

Participants: Angeliki Kritikakou.

Recently, Network-on-Chip (NoC)-based Multi-Processor System-on-Chips (MPSoCs) have become popular computing platforms for real-time applications due to high communication performance and energy efficiency over traditional bus-based MPSoCs. Due to the nature of network structures, network congestion along with transient faults, can significantly affect communication efficiency and system reliability. Most existing works have rarely focused on the concurrent optimization of network contention, reliability, and energy consumption. We study the problem of contention and reliability-aware task mapping under real-time constraints for dynamic voltage and frequency scaling-enabled NoC [30]. The problem entails optimizing voltage/frequency on cores and links to reduce energy consumption and ensure system reliability, while task mapping and slack time are adopted to alleviate network contention and reduce latency. We aim to minimize computation and communication energy and balance workload. This problem is formulated as a mixed-integer nonlinear programming, and we present an effective linearization scheme that equivalently transforms it into a mixed-integer linear programming to find the optimal solution. To reduce computation time, we propose a three-step heuristic, including task allocation, frequency scaling and edge scheduling, and communication contention management. Finally, we perform extensive simulations to evaluate the proposed method. The results show we can achieve 31.6% and 21.7% energy savings, with 95,5% and 98.6% less contention than the existing methods.

7.10 Side-Channel Attacks on Embedded Artificial Intelligence

Participants: Benoit Coqueret, Guillaume Lomet, Olivier Sentieys.

Artificial intelligence, and specifically DNNs, has rapidly emerged in the past decade as the standard for several tasks from specific advertising to object detection. The performance offered has led DNN algorithms to become a part of critical embedded systems, requiring both efficiency and reliability.

In particular, DNNs are subject to malicious examples designed in a way to fool the network while being undetectable to the human observer: the adversarial examples. While previous studies propose frameworks to implement such attacks in black box settings, those often rely on the hypothesis that the attacker has access to the logits of the neural network, breaking the assumption of the traditional black box. In [33], we investigate a real black box scenario where the attacker has no access to the logits. In particular, we propose an architecture-agnostic attack which solve this constraint by extracting the logits. Our method combines hardware and software attacks, by performing a side-channel attack that exploits electromagnetic leakages to extract the logits for a given input, allowing an attacker to estimate the gradients and produce state-of-the-art adversarial examples to fool the targeted neural network. Through this example of adversarial attack, we demonstrate the effectiveness of logits extraction using side-channel as a first step for more general attack frameworks requiring either the logits or the confidence scores.

Dataflow neural network accelerators efficiently process AI tasks, and cloud-based FPGAs, along with ready-to-use frameworks and pre-trained models, simplify their deployment. However, this convenience makes them vulnerable to malicious actors seeking to reverse engineer valuable Intellectual Property (IP) through Side-Channel Attacks (SCA). This year, we proposed a methodology to recover the configuration of a dataflow accelerator generated with the FINN framework. By using unsupervised dimensionality reduction, we reduce computational overhead, enabling lightweight classifiers to recover both the folding and quantization parameters. Our experiments demonstrate that the attack phase requires only 2 ms to recover over 95% of these parameters for a FINN-based accelerator running a CNN, using a k-NN classifier from two averaged side-channel traces, even with the accelerator dataflow fully loaded. This approach offers a more realistic attack scenario than existing methods.

7.11 Hardware Security

Participants: Amélie Marotta, Ronan Lashermes, Olivier Sentieys.

Cache Side Channel Attacks (CSCA) have been haunting most processor architectures for decades now. Existing approaches to mitigation of such attacks have certain drawbacks namely software mishandling, performance overhead, low throughput due to false alarms, etc. Hence, *"mitigation only when detected"* should be the approach to minimize the effects of such drawbacks. In [21], we propose a novel methodology of fine-grained detection of timing-based CSCA using a hardware-based detection module. The detection results are checked under different workload conditions with respect to the number of attackers, the number of victims having RSA,AES and ECC based encryption schemes like ECIES, and on benchmark applications like MiBench and Embench. More than 98% detection accuracy within 2% of the beginning of an attack can be achieved with negligible false alarms. The detection module has an area and power overhead of 0.9% to 2% and 1% to 2.1% for the targeted RISC-V processor core without cache for 1 to 5 counters, respectively. The detection module does not affect the processor critical path and hence has no impact on its maximum operating frequency. This work is conducted in collaboration with IIT Goa, India.

In the realm of fault injection, Electro-Magnetic Fault Injection (EMFI) attacks have garnered significant attention, particularly for their effectiveness against embedded systems with minimal setup. These attacks exploit vulnerabilities with ease, underscoring the importance of comprehensively understanding EMFI. Recent studies have highlighted the impact of EMFI on PLL, uncovering specific clock glitches that induce faults. However, these studies lack a detailed explanation of how these glitches translate into a specific fault model. Addressing this gap, our research investigates the physical fault model of synchronous clock glitch (SCG), a clock glitch injection mechanism likely to arise from EMFI interactions within the clock network [51]. Through an integrated approach combining experimental and simulation techniques, we critically analyze the adequacy of existing fault models, such as the Timing Fault Model and the Sampling Fault Model, in explaining SCG. Our findings reveal specific failure modes in DFFs, contributing to a deeper understanding of EMFI effects and aiding in the development of more robust defensive strategies against such attacks.

7.12 Computational Memories

Participants: Léo de la Fuente, Olivier Sentieys.

In the Computing-In-Memory (CIM) approach, computations are directly performed within the data storage unit, which often results in energy reduction. This makes it particularly well fitted for embedded systems, highly constrained in energy efficiency. It is commonly admitted that this energy reduction comes from less data transfers between the CPU and the main memory. Nevertheless, preparing and sending instructions to the computational memory also consumes energy and time, hence limiting overall performance. In [41], we present a hardware instruction generation mechanism integrated in computational memories and evaluate its benefit for Integer General Matrix Multiplication (IGeMM) operations. The proposed mechanism is implemented in the computational memory controller and translates macro-instructions into corresponding micro-instructions needed to execute the kernel on stored data. We modified an existing near-memory computing architecture and extracted corresponding energy consumption figures using post-layout simulations for the complete SoC. Our proposed architecture, NEar memory computing Macro-Instruction Kernel Accelerator (NeMIKA), provides an 8.2× speed-up and a 4.6× energy consumption reduction compared to a state-of-the-art CIM accelerator based on micro-instructions, while inducing an area overhead of only 0.1%. This work is conducted in collaboration with CEA List, Grenoble.

7.13 Fault-Tolerant Networks-on-Chip

Participants: Wilfread Guilleme, Hamza Amara, Ibrahim Krayem, Angeliki Kritikakou, Cédric Killian, Emmanuel Casseau, Daniel Chillet.

Network-on-Chip (NoC) has emerged as the primary interconnect solution in the multicore and manycore era since the early 2010s. However, the reduction in transistor size has increased the sensitivity of these systems to faults. Meanwhile, artificial intelligence (AI) algorithms are increasingly being integrated across various application domains, especially in embedded systems, to facilitate edge computing and minimize data transfer to the cloud.

For embedded neural networks (NNs), faults like Single-Event Upsets (SEUs) can significantly affect their reliability. To address this challenge, previous works have been done about SEU layers sensitivity of AI models. On previous study, we demonstrated that faults causing bit flips from 0 to 1 significantly impact classification outcomes and we proposed an HTAG (Hardening Technique using And Gates) to mitigate these faults. This technique relies on bit duplication, and without performing error correction, it ensures that the bit value is set to 0 when the two duplicated bit instances differ. While this technique may be effective for floating-point number representation, it may not be suitable for fixed-point numbers. Given that fixed-point representation is commonly used in hardware implementations, especially in FPGA designs, we introduced a new technique called VANDOR. This method identifies the sign of the numbers and applies adapted mitigation strategies to minimize the impact of faults as much as possible [46] [45].

Additionally, since faults can result from intentional attacks on the NoC, we also analyzed their impact on data transfer when a specific compression format is applied to NoC traffic. Using image transfer as a case study, we examined the degradations caused by attacks on the payload of data traffic for both uncompressed data and FlitZip-compressed data [81]. Experimental results indicate that the mean square error (MSE) of uncompressed data increases more rapidly with higher fault injection rates compared to the FlitZip-based approach. However, FlitZip-based approach also suffers from the potential presence of hardware Trojans (HT) within some of the intellectual properties (IP) embedded in the NoC. To tackle this issue, we developed a lightweight mitigation technique that focuses on protecting the bases information by utilizing unused bits in the header flit. Experimental results demonstrate that our approach effectively reduces the impact of hardware Trojan attacks on compressed packet data, achieving a mean squared error reduction of up to 67% and a loss in compression ratio of around 8% only when tested on the CIFAR-10 dataset [35].

8 Bilateral contracts and grants with industry

8.1 Bilateral Contracts with Industry

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

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

8.2 Bilateral Grants with Industry

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

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

Participants: Olivier Sentieys, Benoit Coqueret.

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

Participants: Daniel Chillet.

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

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

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

Participants: Bertrand Le Gal, Simon Rokicki, Olivier Sentieys.

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

KeySom SAS is also funding the PhD thesis of Anis Yagoub on the exploration of dynamically reconfigurable floating-point units for transprecision computation in deep learning.

8.3 Informal Collaborations with Industry

Participants: Olivier Sentieys, Silviu-Ioan Filip.

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

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 Inria associate team

AxTRADE

Title: Approximation-aware Training of DNNs for Edge AI Hardware

Duration: 2024 - 2026

Coordinator: Gayathri Ananthanarayanan (gayathri@iitdh.ac.in)

Partners:

- Indian Institute of Technology Dharwad (Inde)

Inria contact: Marcello Traiola

Summary: In the context of smart and reconfigurable edge AI hardware platforms, e.g. smart cameras, there is a significant demand of configurable computation effort within the system platform. This need arises from the limited energy available in such edge platform. The quality of input data to such systems changes based on scene-dependent factors such as lighting conditions, environmental shifts, and scene complexity. Hence the electronic platform has to offer configurable computation quality levels through Approximate Computing (AxC) and at the same time avoid critical accuracy drops by adapting the DNN model to the different computation quality levels available. This SW/HW synergy has the potential to lead to high energy savings without compromising the accuracy of the output.

Choosing the quality levels that the system has to offer is a complex task, as it involves considerations like reconfiguration time, memory usage, storage requirements, energy consumption, quality of service, and performance guarantees. In resource-constrained devices it is crucial to minimize the overhead associated with the reconfiguration capabilities. In fact, generating, retraining, and storing inside the edge device all possible variants of a DNN model to adapt to various approximations in the system is impractical and costly. Therefore, in this collaborative project we propose to investigate, design, and implement smart approaches to generate an optimal number of DNN variants that adapts to different level of accuracy while minimizing the associated hardware overheads. This must be achieved without compromising runtime performance and output accuracy.

EdgeTrain

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

Duration: 2022 - 2024

Coordinator: Guy Lemieux (lemieux@ece.ubc.ca)

Partners:

- University of British Columbia Vancouver (Canada)

Inria contact: Silviu-Ioan Filip

Summary: For many application scenarios, such as autonomous driving and healthcare wearables, there is a strong need for real-time and on-site learning of deep neural network (DNN) in order to enable them to proactively learn from new data and adapt to changing environments. Compared to cloud-based (re)training, training locally helps to avoid costly data transfers between data centers and edge devices, and to reduce communication cost/latency and offer enhanced privacy. Much work on accelerating DNN training has focused on resource-rich distributed computing scenarios and a large mini-batch regime. While such approaches definitely reduce training time, they increase energy consumption significantly. Reducing the numerical precision of basic arithmetic operations, on the other hand, is a general way to increase performance and energy efficiency in computing, as hardware energy efficiency usually improves quadratically with the decrease in bit precision.

The research direction that our project aims to address is the design and development of an automated process for hardware training on edge devices that is tailored to a specific neural network architecture. The main challenge in this setting is how to reduce the hardware complexity of the required operators (at the arithmetic level) such that the training process is sure to converge. We will in particular explore the use of several precision levels during the training process, with the goal of using the lowest numerical precision possible for as much of the training process as possible.

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

9.1.2 Participation in other International Programs

IntelliVIS

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

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

Partner Institution: IIT Goa (India)

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

LRS

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

Title: Loop unRolling Stones: compiling in the polyhedral model

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

Inria contact: Steven Derrien

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

Informal International Partners

- Dept. of Electrical and Computer Engineering, Concordia University (Canada), Optical network-on-chip, manycore architectures.
- LSSI laboratory, Québec University in Trois-Rivières (Canada), Design of architectures for digital filters and mobile communications.

- University of Trento (Italy), Reliability analysis and radiation experiments
- School of Informatics, Aristotle University of Thessaloniki (Greece), Memory management, fault tolerance
- Karlsruhe Institute of Technology - KIT (Germany), Loop parallelization and compilation techniques for embedded multicores.
- PARC Lab., Department of Electrical, Computer, and Software Engineering, the University of Auckland (New-Zealand), Fault-tolerant task scheduling onto multicore.
- Ruhr - University of Bochum - RUB (Germany), Reconfigurable architectures.
- School of Automation, Southeast University (China), Fault-tolerant task scheduling onto multi-core.
- Shantou University (China), Runtime efficient algorithms for subgraph enumeration.
- University of Science and Technology of Hanoi (Vietnam), Participation in the Bachelor and Master ICT degrees.
- Department of Electrical and Computer Engineering, University of Naples (Italy), Digital Hardware Design Space Exploration for Approximate-Computing-based Applications
- Department of Control and Computer Engineering, Politecnico di Torino (Italy), Fault tolerance of Deep Neural Network hardware accelerators
- Department of Computer Science, University of Verona (Italy), Assertion-driven Design Exploration of Approximate Hardware

9.2 International research visitors

Alessio Colucci, PhD student at TU Wien, Austria, visited TARAN from April 2024 until May 2024.

9.3 European initiatives

9.3.1 Horizon Europe

EDF-EU ARCHYTAS

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

- Program: EDF-2023-RA
- Project acronym: ARCHYTAS
- Project title: ARCHitectures based on unconventional accelerators for dependable/energY efficienT AI Systems
- Duration: January 2025 – December 2027
- Coordinator: Alessandro MORANDO, Iveco Defence Vehicles, Italy
- Other partners: 5 industrial LE and midcap partners (IDV, IWATT, ITECH, SENER and STM), 6 SMEs (NUREN, SPINV, UBOTICA, BRIGHT, MR and UPMEM), and 14 RTOs (UNITN, POLIMI, UNIBO, AUTH, FHG_IPMS, JMU, CSIC, SU with affiliate CNRS, NAMLAB, UG, NTNU, and UNIVREN with affiliate INRIA (TARAN)

The ARCHYTAS project aims to investigate and study the feasibility of non-conventional AI accelerators for defence applications that take advantage of novel technologies at the device and package level: optoelectronic-based accelerators, volatile and non-volatile processing-in-memory, and neuromorphic devices. The project will also investigate the integration of CMOS-based systems with analog accelerator devices and their organisation by integrating them in a multi-chip (chiplet) configuration. Moreover, ARCHYTAS will investigate new programming models to improve the programmability, performance portability and in general productivity of newer emerging parallel systems by following a HW-AI co-design. The targeted gains the ARCHYTAS AI accelerators will be measured and validated within the context of defence AI use cases that seek leverage the efficiency and tactical gains for military missions through use of computer vision and increased autonomy of defence assets in land, aerial, maritime and space settings. The innovations developed in ARCHYTAS distinctly addresses the challenges encountered in the use cases by presenting solutions optimized for energy consumption, speed, and cost. The technological ambition of ARCHYTAS is to bridge the gaps in multi-modal sensing integration and AI processing, providing solutions that will fit with the nonfunctional requirements of future autonomous vehicles for defence applications. These innovations hold the potential for disruption in the defence domain by setting new benchmarks for performance and efficiency. Quantitatively, the aim of the use case owners is to achieve transformative gains in AI processing speed and energy efficiency, targeting improvements of several orders of magnitude over existing solutions to enhance significantly European defence capabilities in different operational domain, particularly where autonomous systems are key areas.

9.3.2 Other european programs/initiatives

Inria/DFKI FAIRe

Participants: Romain Facq, Silviu Filip, Olivier Sentieys.

- Program: Inria/DFKI
- Project acronym: FAIRe
- Project title: Frugal Artificial Intelligence in Resource-limited environments
- Duration: Mar 2024 – Dec 2028
- Coordinator: Silviu Filip, Taran, Christoph Lüth, DFKI
- Other partners: Taran, Cash, Corse, DFKI CPS, DFKI AV, DFKI ASR, DFKI RIC

Artificial intelligence (AI) is finding many new applications in the physical world. For this, AI applications need to run on embedded, cyber-physical devices with limited resources and less than ideal conditions. We call this frugal AI — AI with a small memory footprint, using less computational power, and working with fewer data. To develop frugal AI applications, FAIRe develops a comprehensive approach on all abstraction layers of an AI application, unifying previously disjoint approaches to this problem: developing special hardware extensions, enabling compiler support to utilize these extensions, and algorithms which cope with resource restrictions by e.g. quantization or continual learning. A case study from the area of domestic robotics covering all these aspects will demonstrate our approach in practice.

9.4 National initiatives

9.4.1 ANR RAKES

Participants: Olivier Sentieys, Cédric Killian, Abhijit Das.

- Program: ANR PRC

- Project acronym: RAKES
- Project title: Radio Killed an Electronic Star: speed-up parallel programming with broadcast communications based on hybrid wireless/wired network on chip
- Duration: June 2019 - June 2024
- Coordinator: TIMA
- Other partners: TIMA, TARAN, Lab-STICC

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

9.4.2 ANR Opticall2

Participants: Olivier Sentieys, Cédric Killian, Daniel Chillet.

- Program: ANR PRCE
- Project acronym: Opticall2
- Project title: on-chip OPTIcal interconnect for ALL to ALL communications
- Duration: Dec. 2018 - May 2024
- Coordinator: INL
- Other partners: INL, TARAN, C2N, CEA-LETI, Kalray

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

9.4.3 ANR SHNOC

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

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

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

9.4.4 ANR FASY

Participants: Angeliki Kritikakou, Marcello Traiola, Olivier Sentieys.

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

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

9.4.5 ANR Re-Trusting

Participants: Olivier Sentieys, Angeliki Kritikakou, Marcello Traiola, Silviu-Ioan Filip.

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

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

9.4.6 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.

9.4.7 ANR LOTR

Participants: Steven Derrien, Simon Rokicki.

- Program: ANR PRC
- Project acronym: LOTR
- Project title: Lord Of The RISCs
- Duration: Oct. 2023 - Sep. 2027

- Coordinator: Steven Derrien
- Other partners: CEA, TARAN, PACAP

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

9.4.8 CYBERPROS

Participants: Olivier Sentieys.

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

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

9.4.9 PEPR ARSENE

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

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

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

9.4.10 ANR RADYAL

Participants: Marcello Traiola, Olivier Sentieys.

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

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

9.4.11 ANR SEC-V

Participants: Bertrand Le Gal.

- Program: ANR PRCE
- Project acronym: SEC-V
- Project title: open-source, secure and high-performance processor core based on the RISC-V ISA
- Duration: Oct 2021 – Apr 2025
- Coordinator: Sebastien Pillement, IETR, Nantes
- Other partners: TARAN, LS2N, THALES TRT, THALES INVIA

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

9.4.12 PEPR HOLIGRAIL

Participants: Nesrine Sfar, Rémi Garcia, Mehdi El Arar, Silviu Filip, Olivier Sentieys.

- Program: PEPR IA

- Project acronym: HOLIGRAIL
- Project title: HOLIistic approaches to GReener model Archi-tectures for Inference and Learning
- Duration: Oct 2023 – Dec 2029
- Coordinator: Olivier Sentieys, Taran
- Other partners: CEA List, INSA Lyon, Inria Corse, Grenoble-INP

Accelerators of artificial intelligence algorithms currently consume much more power than they should, in particular in the learning phase. The many aspects of this question are too often considered in isolation. Based on the complementary expertise of the partners, and thanks to the integration into the rich community build by the PEPR on foundation of frugal AI, we will instead systematically look at a holistic, global comprehension of all these issues in established and upcoming AI algorithms. We will therefore combine more compact and efficient number representations, hard-ware-aware training algorithms that enhance structured sparsity, coding compactness and tensor transformations, with their adaptation to efficient hardware mechanisms and compiler optimizations. Our ambition is to provide breakthroughs in efficiency when running inference and training algorithms on specialized hardware. The results are intended to be integrated into development solutions for embedded systems, in particular within the DeepGreen national platform for the deployment of deep learning in embedded systems.

9.4.13 PEPR ARCHI-SESAM

Participants: Marcello Traiola, Olivier Sentieys.

- Program: PEPR Cloud
- Project acronym: ARCHI-SESAM
- Project title: Converged, Efficient and Safe Architecture based on Near Memory Accelerators
- Duration: Oct 2023 – Dec 2029
- Coordinator: Denis Dutoit, CEA, Grenoble
- Other partners: Taran, CEA List, IMT, Inria Convecs, Grenoble-INP

European sovereignty in the cloud also means sovereignty over hardware, especially processors and accelerators. Improvement of processor performance requires hardware architectures that evolve towards more parallelism (multi-core), more specialization (accelerators), a closer relationship between computing and memory and new types of interconnections between components. On the other hand, by dissociating hardware resources (computing, memory, interconnection) from logical resources, virtualization facilitates the deployment of converged architectures that bring together the computing, storage and network infrastructure. The cloud gains in modularity, speed and agility for the deployment of new services with optimal use of resources. Hardware disaggregation on the one hand and resource virtualization on the other are making the intermediate adaptation layer increasingly complex, difficult to validate and prone to failure. The Archi-CESAM project proposes to rethink the hardware (computing, memory and interconnection) so that it is co-designed with the application in a perspective of converged architecture and trust, in an environment known for its abundance of data to be processed. The Archi-CESAM project addresses this major evolution of the Cloud in a global and coordinated approach between distributed architectures, acceleration, interconnection and security bricks, without forgetting the design methods.

9.4.14 Inria Challenge CocoRISCo

Participants: Simon Rokicki, Ronan Lashermes, Olivier Sentieys.

- Program: Inria Challenge
- Project acronym: CocoRISCo
- Project title: Hardware-software interface for general purpose computing with RISC-V
- Duration: Oct 2024 – Sep 2028
- Coordinator: Olivier Sentieys, Taran, Arthur Pérais, TIMA
- Other partners: CEA List, Inria (Corse, Benagil, Pacap, Sushi, Madmax, Taran)

CocoRISCo focuses on the hardware and low-level software aspects of computer systems. Specifically, those systems have dramatically evolved in the past decades, yet many interfaces between hardware and software layers have remained in place with little changes. Indeed, hardware has become heavily multithreaded (e.g., multi-, many-cores, GPUs), heterogeneous (e.g., dedicated accelerators attached to CPUs), and open to vulnerabilities caused by increased complexity (e.g., speculation-based attacks à la Spectre). We aim to leverage the RISC-V open Instruction Set Architecture (ISA) – the interface between software and the CPU – to revisit and improve those aspects, for instance by exposing more hardware features to the programmer. CocoRISCo will gather 5 Inria teams that have a background in architecture, microarchitecture, compilation, operating systems, and security, along with the SLS team of the TIMA laboratory and the DSCIN of laboratory CEA List.

9.4.15 RAPID FOCH

Participants: Joseph Paturel, Olivier Sentieys.

- Program: RAPID
- Project acronym: FOCH
- Project title: Development of a fault-tolerant FPGA with tests in high-radiation environments
- Duration: Jan 2023 – Dec 2025
- Coordinator: NanoXplore
- Other partners: Taran, Onera, Nucletudes

FPGA components are widely used in aerospace and military applications. This project aims to consider constrained radiative environments and to develop a fault-tolerant FPGA IP. A RISC-V processor will be used as a test case for implementation on the FPGA IP and for evaluation in high-radiation environments.

9.4.16 Inria Challenge OmicFinder

Participants: Bertrand Le Gal, Olivier Sentieys.

- Program: Inria Challenge
- Project acronym: OmicFinder

- Project title: Biological data indexation
- Duration: Oct 2023 – Dec 2027
- Coordinator: Pierre Peterlongo, Genscale
- Other partners: Taran, Dyliss, Zenith, CEA-GenoScope, Elixir, Pasteur Institute, CEA-CNRGH, Mediterranean Institute of Oceanography

Genomic data enable critical advances in medicine, ecology, ocean monitoring, and agronomy. Precious sequencing data accumulate exponentially in public genomic data banks such as the ENA. A major limitation is that it is impossible to query these entire data (petabytes of sequences). OmicFinder aims to provide a novel global search engine making it possible to query nucleotidic sequences against the vast amount of publicly available genomic data. The central algorithmic idea of a genomic search engine is to index and query small exact words (hundreds of billions over millions of datasets), as well as the associated metadata. In addition to the creation of fundamental novel algorithms and data structures, the project develops new approaches to improve the query experience and the answer information by integrating the Semantic Web technologies framework. In view of the considered volume of data, a part of the research focuses on clever index distribution. Throughout the project, we are committed to proposing methods that minimize the environmental impact generated by the massive use of the tools that will be produced, in particular through the use of specialized hardware.

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- M. Traiola was the General co-chair of IEEE IOLTS 2024.
- A. Kritikakou was the General co-chair of IEEE IOLTS 2024.

Member of the organizing committees

- D. Chillet was in the Organizing Committee of Rapido 2024.
- O. Sentieys and M. Traiola were members of the IEEE/ACM DATE Executive Committee, 2024.
- M. Traiola was the Review and Programme Operations Co-Chair of IEEE/ACM DATE 2024.
- F. Fernandes dos Santos was the Publication Chair and Media Chair of IEEE IOLTS 2024.
- A. Kritikakou was the Education chair IEEE ESWEEK 2024, PhD Forum chair ETS2 024, and Academic jury of McCluskey Doctoral Thesis Award ITC 2024.

10.1.2 Scientific events: selection

Chair of conference program committees

- O. Sentieys was co-chair of the Focus Sessions at IEEE/ACM DATE Executive Committee, 2024.

Member of the conference program committees

- D. Chillet was member of the technical program committee of HiPEAC Rapido, HiPEAC WRC, DSD, ComPAS, DASIP, ARC.
- S. Filip was member of the technical program committee for IEEE ARITH, IEEE ASAP.
- M. Traiola was member of the technical program committee for IEEE ICCAD, IEEE VTS, IEEE ETS, IEEE IOLTS, IEEE DFT, IEEE/ITRI VLSI-DAT, ACM CF, IEEE eARTS workshops, Approximate Computing (AxC) workshop.
- F. Fernandes dos Santos was a member of the technical program committee for the IEEE IOLTS, IEEE DFT, CASES, RADECS, and IEEE DATE.
- A. Kritikakou was member of the technical program committee of DAC, DAC-LBR, DATE, EMSOFT, EMSOFT-LBR, CASES, ISVLSI, ETS, SAMOS, DS-RT, AI-TREATS, COMPAS.
- O. Sentieys served as a committee member in the IEEE EDAA Outstanding Dissertations Award (ODA) 2023.
- O. Sentieys was a member of technical program committee of IEEE/ACM ICCAD, IEEE FPL, ACM ENSSys, ACM SBCCI, ARC.
- S. Rokicki is a member of program committee of IEEE/ACM DATE and IEEE/ACM ESWEEK/CASES

10.1.3 Journal

Member of the editorial boards

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

Reviewer - reviewing activities

- M. Traiola was a reviewer for IEEE (ToC, TCAD, TECT, TCA, TDMR, TNS, Design&Test) and ACM journals (TECS, JETC, JATS, TODAES)
- F. Fernandes dos Santos was a reviewer for IEEE (TNS, TC, TAES, JETCAS, TECS) and ACM/Springer journals (JSA, JSC, Mic. Reliability, and Microprocessors and Microsystems).
- D. Chillet was a reviewer for Microprocessors and Microsystems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Journal of Systems Architecture, and ACM Transactions on Architecture and Code Optimization.
- B. Le Gal was reviewer for IEEE Wireless Communications Letters, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Transactions on Circuits and Systems II: Express Briefs, IEEE Signal Processing Letters, IEEE Communications Letters, IEEE Signal Processing Letters, and IEEE Transactions on Image Processing.
- O. Sentieys was a reviewer for IEEE Transactions on VLSI Systems and ACM Transactions on Embedded Computing Systems.
- A. Kritikakou was a reviewer for IEEE (D&T, TC, TPDS, TCAD, TECT, etc) and ACM journals (TECS, JETC, CS, TODAES etc)

10.1.4 Invited talks

- O. Sentieys gave an invited talk on Design and Exploration of RISC-V Cores from High-Level Specifications at the 2nd Sino-European RISC-V Workshop, Hong-Kong, 27-29 Nov. 2024.

10.1.5 Leadership within the scientific community

- D. Chillet is a member of the French National University Council in Signal Processing and Electronics (CNU - Conseil National des Universites, 61ème section) since 2019.
- D. Chillet is member of the Board of Directors of Gretsi Association.
- A. Kritikakou is a member of the French National University Council in Computer Science (CNU - Conseil National des Universites, 27ème section) since 2022.
- A. Kritikakou is co-animator of the "High performance embedded computing" topic of GDR SoC2.
- O. Sentieys is a member of the steering committee of a CNRS Spring School for graduate students on embedded systems architectures and associated design tools (ARCHI).
- O. Sentieys is a member of the steering committee of GDR SoC2.

10.1.6 Scientific expertise

- S. Derrien was a member of the ANR Scientific Evaluation Committee CE25 "Software science and engineering - Multi-purpose communication networks, high-performance infrastructure".
- D. Chillet has done a scientific expertise for a research project submitted to German Research Foundation.

10.1.7 Research administration

- S. Derrien was the head of the D3 "Computer Architecture" Department of IRISA Lab until Aug. 2024.

10.1.8 Standardization activities

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

10.2 Teaching - Supervision - Juries

10.2.1 Teaching administration

- A. Kritikakou is a member of the Examination Committee of Industrial Engineering Sciences and Computer Engineering (SII) Aggregation.
- E. Casseau is in charge of the Department of “Digital Systems” at ENSSAT Engineering Graduate School.
- B. Le Gal is associate director of studies at ENSSAT Engineering Graduate School since Nov 2024.
- D. Chillet was associate director of studies at ENSSAT Engineering Graduate School until Sept 2024.
- S. Rokicki is responsible of the second year of the computer science department at ENS Rennes

10.2.2 Teaching

- D. Chillet: embedded processor architecture, 20h, Enssat (M1)
- D. Chillet: multimedia processor architectures, 30h, Enssat (M2)
- D. Chillet: advanced processor architectures, 20h, Enssat (M2)
- D. Chillet: advanced processor architectures, 15h, Embedded Systems Master (M2)
- D. Chillet: micro-controller, 32h, Enssat (L3)
- D. Chillet: low-power digital CMOS circuits, 4h, UBO (M2)

- A. Kritikakou: Tools and programming in C, 24.75h, istic (L3)
- A. Kritikakou: Computer programming, 22.5h, istic (L3)
- A. Kritikakou: Unix commands and programming, 6.75h, istic (L3)
- A. Kritikakou: Fault tolerant embedded systems, 6h, INSA (M2)
- A. Kritikakou: Energy sobriety of digital architectures, 7.5h, INSA (M2)
- B. Le Gal: Digital fundamentals, 24h, ENSSAT (L3)
- B. Le Gal: VHDL design, 32h, ENSSAT (M1)
- B. Le Gal: Hardware & software verification, 12h, ENSSAT (M1)
- B. Le Gal: Processor design (RISC-V), 26h, ENSSAT (M1)
- B. Le Gal: Real-time programming, 26h, ENSSAT (M1)
- B. Le Gal: Software compilation, 16h, ENSSAT (M2)
- B. Le Gal: System on Chip design, 18h, ENSSAT (M2)
- B. Le Gal: High performance computing, 16h, ENSSAT (M2)
- S. Rokicki: Compilers, 24h, ENS Rennes
- S. Rokicki: Advanced Compilers, 10h, ENS Rennes
- O. Sentieys: Hardware Accelerators for Deep Neural Networks, 54h, Master of Embedded Systems, ISTIC (M2)
- O. Sentieys, High-Level synthesis, 20h, Master of Computer Science, ISTIC (M2)
- M. Traiola: Operating Systems, 24h, ENS Rennes (Aggregation Mecatronique)
- M. Traiola: Hardware Accelerators for Deep Neural Networks, 12h TP, Master of Embedded Systems, ISTIC (M2)
- F. Fernandes dos Santos: TinyML at Master SE ISTIC(6h).
- F. Fernandes dos Santos: C/Unix for L3 (39h TPs)

10.2.3 Educational class

- M. Traiola gave a 2-hour educational class on "Efficient Neural Networks: from SW optimization to specialized HW accelerators" at Embedded System Week 2024 [34]

10.2.4 PhD Supervision

- PhD defended: Jean-Michel Gorius, High-Level Synthesis of Instruction Set Processors, Dec. 2024, S. Derrien, S. Rokicki.
- PhD defended: Corentin Ferry, CAutomating the derivation of memory allocations for acceleration of polyhedral programs, Feb. 2024, S. Derrien, T. Yuki and S. Rajopadhye (co-tutelle between Univ Rennes and Colorado State University).
- PhD defended: Cédric Gernigon, Neural Network Quantization Methods for FPGA On-Board Processing of Satellite Image, Dec. 2024, O. Sentieys, S. Filip.
- PhD defended: Ibrahim Krayem, Methods for fast exploration of manycore architectures based network-on-chip with emerging technologies, Feb. 2024, C. Killian, D. Chillet.
- PhD defended: Leo Pradels, Efficient CNN Inference Acceleration on FPGAs: A Pattern Pruning-Driven Approach, CIFRE Thesis with Safran, Dec. 2024, D. Chillet, S. Filip.
- PhD defended: Louis Narmour, Optimizations of Polyhedral Reductions and their use in Algorithm-Based Fault Tolerance, Dec. 2024, S. Derrien, T. Yuki and S. Rajopadhye (co-tutelle between Université de Rennes and Colorado State University).

- PhD in progress: Sohaib Errabii, Dynamically Configurable Deep Neural Network Hardware Accelerators, April 2024, M. Traiola, O. Sentieys.
- PhD in progress: Anis Yagoub, Exploring dynamically reconfigurable floating-point units for transprecision computation in deep learning, CIFRE Thesis with Keysom SAS, Sep. 2024, B. Le Gal, O. Sentieys.
- PhD in progress: Lucas Roquet, Dependability Evaluation and Enhancing Methods for Large Machine Learning Models, Oct. 2024, F. Fernandes dos Santos, A. Kritikakou.
- PhD in progress: Romain Facq, Exploring low precision arithmetic for continual learning tasks on edge devices, Oct. 2024, O. Sentieys, S. Filip.
- PhD in progress: Nesrine Sfar, Compression, fine-tuning, and hardware acceleration of Transformer-based models, Dec. 2024, O. Sentieys, S. Filip.
- PhD in progress: Dylan Leothaud, Automatic synthesis of secure and predictable processors for the Internet of Thing, Oct. 2023, S. Derrien, S. Rokicki.
- PhD in progress: Leo Pajot, Soft-core processor with dynamic binary execution exploiting instruction-level parallelism, CIFRE Thesis with Keysom SAS, Sep. 2023, B. Le Gal, S. Rokicki.
- PhD in progress: Oussama Ait Sidi Ali, Virtualisation of a multi-mission telemetry receiver, CIFRE Thesis with Safran, Apr. 2022, B. Le Gal.
- PhD in progress: Hamza Amara, Detection and countermeasures for DoS attack in Noc-based SoC using machine learning, Oct. 2022, E. Casseau, D. Chillet, C. Killian.
- PhD in progress: Herinomena Andrianatrehina, Ensuring confidentiality in modern Out-of-Order cores, Nov 2022, S. Rokicki, R. Lashermes.
- PhD in progress: Gaetan Barret, Predictive model of energy consumption of cloud-native applications, Nov. 2022, D. Chillet.
- PhD in progress: Sami Ben Ali, Efficient Low-Precision Training for Deep Learning Accelerators, Jan. 2022, O. Sentieys, S. Filip.
- PhD in progress: Benoit Coqueret, Physical Security Attacks Against Artificial Intelligence Based Algorithms, CIFRE Thesis with Thales, Nov. 2022, O. Sentieys, M. Carbone (Thales), G. Zaid (Thales).
- PhD in progress: Wilfread Guilleme, Fault Tolerant Hardware Architectures for Artificial Intelligence, Oct. 2022, D. Chillet, C. Killian, A.Kritikakou.
- 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: Romaric (Pegdwende) Nikiema, Time-guaranteed and reliable execution for real-time multicore architectures, Oct. 2022, A. Kritikakou, M. Traiola
- 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.
- PhD in progress: Léo De La Fuente, In-Memory Computing for Ultra Low Power Architectures, Nov. 2021, O. Sentieys, J.-F. Christmann (CEA).
- PhD in progress: Seungah Lee, Efficient Designs of On-Board Heterogeneous Embedded Systems for Space Applications, Nov. 2021, A. Kritikakou, E. Casseau, R. Salvador, O. Sentieys.
- PhD in progress: Amélie Marotta, Emp-error: EMFI-Resilient RISC-V Processor, Oct. 2021, O. Sentieys, R. Lashermes (LHS), Rachid Dafali (DGA).

10.3 Popularization

The Smolphone project is a collaborative initiative with M. Quinson from the Inria Magellan team, aimed at rethinking the development of a frugal smartphone. The goal is to explore modifications to hardware, software, and feature sets to significantly extend the device's lifecycle. The project received initial funding through an Inria AEx grant, enabling the first stages of development. In parallel, students contributed by designing a heterogeneous system combining a CPU and an MCU, focusing on reducing energy consumption as a key objective [77].

In December 2024, Olivier Sentieys was invited for a video interview about AI compression and acceleration in the famous website L'esprit sorcier, an educational medium for the popularisation of science. The video is still being edited, and should be released in early 2025.

Members of TARAN participate to the working group at IRISA/Inria on reducing GHG emissions from business travel [70].

11 Scientific production

11.1 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 (cit. on p. 11).

[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 (cit. on p. 7).

[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 (cit. on p. 6).

[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 (cit. on p. 7).

[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 (cit. on p. 7).

[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 (cit. on p. 7).

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

11.2 Publications of the year

International journals

[21] P. Bhade, J. Paturel, O. Sentieys and S. Sinha. 'Lightweight Hardware-Based Cache Side-Channel Attack Detection for Edge Devices (Edge-CaSCADe)'. In: *ACM Transactions on Embedded Computing Systems (TECS)* 23 (10th June 2024), pp. 1–27. DOI: 10.1145/3663673. URL: https://hal.science/hal-04764627 (cit. on p. 17).

[22] P. R. Bodmann, M. Saveriano, A. Kritikakou and P. Rech. 'Neutrons Sensitivity of Deep Reinforcement Learning Policies on EdgeAI Accelerators'. In: *IEEE Transactions on Nuclear Science* 71.8 (Aug. 2024), pp. 1480–1486. DOI: 10.1109/TNS.2024.3387087. URL: https://hal.science/hal-04731813 (cit. on p. 15).

[23] A. Bosio, S. Germiniani, G. Pravadelli and M. Traiola. ‘Syntactic and Semantic Analysis of Temporal Assertions to Support the Approximation of RTL Designs’. In: *Journal of Electronic Testing: : Theory and Applications* 40.2 (23rd Apr. 2024), pp. 199–214. DOI: 10.1007/s10836-024-06115-9. URL: https://hal.science/hal-04726657 (cit. on p. 14).

[24] F. Fernandes dos Santos, L. Carro, F. Vella and P. Rech. ‘Assessing the Impact of Compiler Optimizations on GPUs Reliability’. In: *ACM Transactions on Architecture and Code Optimization* 21.2 (12th Jan. 2024), pp. 1–22. DOI: 10.1145/3638249. URL: https://hal.science/hal-04398273 (cit. on p. 14).

[25] F. Fernandes dos Santos, N. Cavagnero, M. Ciccone, G. Averta, A. Kritikakou, O. Sentieys, P. Rech and T. Tommasi. ‘Improving Deep Neural Network Reliability via Transient-Fault-Aware Design and Training’. In: *IEEE Transactions on Emerging Topics in Computing* (2024), pp. 1–12. URL: https://hal.science/hal-04818068. In press (cit. on p. 15).

[26] F. Fernandes dos Santos and P. Rech. ‘Can GPU Performance Increase Faster Than the Code Error Rate?’ In: *Journal of Supercomputing* (20th Apr. 2024), pp. 1–32. DOI: 10.1007/s11227-024-06119-4. URL: https://hal.science/hal-04528798. In press (cit. on p. 14).

[27] F. Fernandes dos Santos and P. Rech. ‘Challenges in Assessing and Improving Deep Neural Networks Reliability’. In: *IEEE Design & Test* (2024), pp. 1–6. URL: https://hal.science/hal-04748638. In press (cit. on p. 15).

[28] P. Kashikar, O. Sentieys and S. Sinha. ‘Combining Weight Approximation, Sharing and Retraining for Neural Network Model Compression’. In: *ACM Transactions on Embedded Computing Systems (TECS)* 23 (11th Sept. 2024), pp. 1–23. DOI: 10.1145/3687466. URL: https://hal.science/hal-04764621 (cit. on p. 13).

[29] B. Loureiro Coelho, F. Fernandes dos Santos, M. Saveriano, G. Allen, A. Daniel, S. Guertin, S. Vartania, E. Wyrwas, C. Frost and P. Rech. ‘Impact of Radiation-Induced Effects on Embedded GPUs Executing Large Machine Learning Models’. In: *IEEE Transactions on Nuclear Science* (2025). URL: https://hal.science/hal-04887365. In press.

[30] L. Mo, X. Li, A. Kritikakou and X. Zhai. ‘Contention and Reliability-Aware Energy Efficiency Task Mapping on NoC-Based MPSoCs’. In: *IEEE Transactions on Reliability* (2024), pp. 1–16. DOI: 10.1109/TR.2024.3377732. URL: https://hal.science/hal-04528715 (cit. on p. 16).

[31] M. Tourres, C. Chavet, B. L. Gal and P. Coussy. ‘Specialized Scalar and SIMD Instructions for Error Correction Codes Decoding on RISC-V Processors’. In: *IEEE Access* 13 (2025), pp. 6964–6976. DOI: 10.1109/ACCESS.2025.3527028. URL: https://hal.science/hal-04891163.

[32] M. Traiola, F. F. dos Santos, P. Rech, C. Cazzaniga, O. Sentieys and A. Kritikakou. ‘Impact of High-Level-Synthesis on Reliability of Artificial Neural Network Hardware Accelerators’. In: *IEEE Transactions on Nuclear Science* (2024), pp. 1–9. DOI: 10.1109/TNS.2024.3377596. URL: https://inria.hal.science/hal-04514579 (cit. on p. 15).

Invited conferences

[33] B. Coqueret, M. Carbone, O. Sentieys and G. Zaid. ‘When Side-Channel Attacks Break the Black-Box Property of Embedded Artificial Intelligence’. In: ESSAI 2024 - 1st edition European Symposium on Security and Artificial Intelligence (ESSAI). Rennes, France, 2024. URL: https://hal.science/hal-04785343 (cit. on p. 16).

[34] M. Traiola, A. Kritikakou, S.-I. Filip and O. Sentieys. ‘Efficient Neural Networks: from SW optimization to specialized HW accelerators’. In: 2024 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES). Raleigh (NC), United States: IEEE; IEEE, 27th Sept. 2024, pp. 1–2. DOI: 10.1109/CASES60062.2024.00009. URL: https://hal.science/hal-04732044 (cit. on p. 34).

International peer-reviewed conferences

[35] H. Amara, C. Killian, D. Chillet and E. Casseau. 'Mitigation of Hardware Trojan in NoC using Delta-Based Compression'. In: SOCC 2024 - 37th IEEE International System-on-Chip Conference. Dresden, Germany: IEEE, 2024, pp. 1–5. DOI: 10.1109/SOCC62300.2024.10737773. URL: https://inria.hal.science/hal-04737447 (cit. on p. 18).

[36] M. Barbareschi, A. Bosio, P. Girard, B. Deveautour, M. Traiola and A. Virazel. 'Approximate Computing for Test and Test of Approximate Computing'. In: IEEE Workshop on Top Picks in VLSI Test and Reliability. San Diego, United States, 2024, pp. 1–2. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-04738424 (cit. on p. 14).

[37] S. Ben Ali, S.-I. Filip and O. Sentieys. 'A Stochastic Rounding-Enabled Low-Precision Floating-Point MAC for DNN Training'. In: DATE 2024 - 27th IEEE/ACM Design, Automation and Test in Europe. Valencia, Spain, 2024, pp. 1–6. URL: https://hal.science/hal-04380270 (cit. on p. 12).

[38] S. Ben Ali, S.-I. Filip, O. Sentieys and G. Lemieux. 'MPTorch-FPGA: a Custom Mixed-Precision Framework for FPGA-based DNN Training'. In: 28th IEEE/ACM Design, Automation and Test in Europe (DATE). Lyon, France, 2025, pp. 1–6. URL: https://hal.science/hal-04882989 (cit. on p. 13).

[39] J. Chen, G. Esposito, F. Fernandes dos Santos, J.-D. Guerrero-Balaguera, A. Kritikakou, M. Krstic, R. Limas, J. E. Rodriguez Condia, M. Sonza Reorda, M. Traiola and A. Veronesi. 'Reliability Assessment of Large DNN Models: Trading Off Performance and Accuracy'. In: VLSI-SoC 2024 - IFIP/IEEE International Conference on Very Large Scale Integration. Tanger, Morocco: IEEE, 2024, pp. 1–10. DOI: 10.5286/ISIS.E.RB2300036). URL: https://hal.science/hal-04736733 (cit. on p. 15).

[40] F. Fernandes dos Santos, M. Traiola and A. Kritikakou. 'Combining Fault Simulation and Beam Data for CNN Error Rate Estimation on RISC-V Commercial Platforms'. In: IOLTS 2024 - 30th IEEE International Symposium on On-Line Testing and Robust System Design. Rennes, France, 2024, pp. 1–8. URL: https://hal.science/hal-04638468 (cit. on p. 15).

[41] L. de la Fuente, J.-F. Christmann, M. Pezzin, M. Remars and O. Sentieys. 'A Hardware Instruction Generation Mechanism for Energy-Efficient Computational Memories'. In: *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. ISCAS 2024 - IEEE International Symposium on Circuits and Systems. ISCAS. Singapour, Singapore: IEEE, 2024. DOI: 10.1109/ISCAS58744.2024.10557870. URL: https://cea.hal.science/cea-04676665 (cit. on p. 17).

[42] C. Gernigon, S.-I. Filip, O. Sentieys, C. Coggiola and M. Bruno. 'AdaQAT: Adaptive Bit-Width Quantization-Aware Training'. In: IEEE 6th International Conference on AI Circuits and Systems (AICAS). Abu Dhabi, United Arab Emirates, 2024. URL: https://hal.science/hal-04549245 (cit. on p. 13).

[43] D. Gnad, J. Krautter, A. Kritikakou, V. Meyers, P. Rech, J. Esteban Rodriguez Condia, A. Ruospo, E. Sanchez, F. Fernandes dos Santos, O. Sentieys, M. B. Tahoori, R. Tessier and M. Traiola. 'Reliability and Security of AI Hardware'. In: ETS 2024 - 29th IEEE European Test Symposium. The Hague, Netherlands: IEEE, 2024, pp. 1–10. URL: https://hal.science/hal-04577494 (cit. on p. 15).

[44] J.-M. Gorius, S. Rokicki and S. Derrien. 'A Unified Memory Dependency Framework for Speculative High-Level Synthesis'. In: CC 2024 - ACM SIGPLAN International Conference on Compiler Construction. Edinburgh (Ecosse), United Kingdom, 2024, pp. 1–13. DOI: 10.1145/3640537.3641581. URL: https://inria.hal.science/hal-04394762 (cit. on p. 11).

[45] W. Guilleme, A. Kritikakou, Y. Helen, C. Killian and D. Chillet. 'HTAG-eNN: Hardening Technique with AND Gates for Embedded Neural Networks'. In: DAC 2024 - IEEE/ACM Design Automation Conference. San Francisco, United States, 2024. DOI: \url{https://dl.acm.org/doi/10.1145/3649329.3657329}. URL: https://inria.hal.science/hal-04689194 (cit. on p. 18).

[46] W. Guillemé, A. Kritikakou, Y. Helen, C. Killian and D. Chillet. 'VANDOR: Mitigating SEUs into Quantized Neural Networks'. In: IOLTS 2024 - IEEE 30th International Symposium on On-Line Testing and Robust System Design. Rennes, France: IEEE, 2024, pp. 1–6. DOI: 10.1109/IOLTS60994.2024.10616081. URL: https://inria.hal.science/hal-04689156 (cit. on p. 18).

[47] M. Hasan Ahmadilivani, A. Bosio, B. Deveautour, F. Fernandes dos Santos, J. David Guerrero Balaguera, M. Jenihhin, A. Kritikakou, R. Limas Sierra, S. Pappalardo, J. Raik, J. Esteban Rodriguez Condia, M. Sonza Reorda, M. Taheri and M. Traiola. 'Special Session: Reliability Assessment Recipes for DNN Accelerators'. In: VTS 2024 - IEEE VLSI Test Symposium. Vol. 10. Tempe AZ USA, United States: IEEE, 2024, pp. 131788–131828. DOI: 10.1109/access.2022.3229767. URL: https://hal.science/hal-04572731 (cit. on p. 15).

[48] L. Iurada, N. Cavagnero, F. Fernandes dos Santos, G. Averta, P. Rech and T. Tommasi. 'Transient Fault Tolerant Semantic Segmentation for Autonomous Driving'. In: UNCV 2024 - 3rd Workshop on Uncertainty Quantification for Computer Vision. Milano, Italy, 2024, pp. 1–6. URL: https://hal.science/hal-04684784 (cit. on p. 15).

[49] S. LEE, E. Casseau, A. Kritikakou, O. Sentieys, R. Salvador and J. Galizzi. 'On-board Payload Data Processing Combined with the Roofline Model for Hardware/Software Design'. In: AeroConf 2024 - IEEE Aerospace Conference. Big Sky, Montana, United States, 2024, pp. 1–12. DOI: 10.1109/AERO58975.2024.10521057. URL: https://inria.hal.science/hal-04423185 (cit. on p. 12).

[50] D. Leothaud, J.-M. Gorius, S. Rokicki and S. Derrien. 'Efficient Design Space Exploration for Dynamic &amp; Speculative High-Level Synthesis'. In: FPL 2024 - 34th International Conference on Field-Programmable Logic and Applications. Turin, Italy: IEEE, 2024, pp. 1–9. DOI: 10.1109/FPL64840.2024.00024. URL: https://hal.science/hal-04615767 (cit. on p. 11).

[51] A. Marotta, R. Lashermes, G. Bouffard, O. Sentieys and R. Dafali. 'Characterizing and Modeling Synchronous Clock-Glitch Fault Injection'. In: COSADE 2024 - Constructive Side-Channel Analysis and Secure Design. Vol. 14595. Lecture Notes in Computer Science. Gardanne, France: Springer Nature Switzerland, 3rd Apr. 2024, pp. 3–21. DOI: 10.1007/978-3-031-57543-3_1. URL: https://inria.hal.science/hal-04549548 (cit. on p. 17).

[52] R. P. Nikiema, M. Traiola and A. Kritikakou. 'Impact of Compiler Optimizations on the Reliability of a RISC-V-based Core'. In: DFT 2024 - 37th IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems. Oxfordshire, United Kingdom, 2024, pp. 1–1. URL: https://hal.science/hal-04731794 (cit. on p. 16).

[53] J. Pottier, T. Nieddu, B. Le Gal, S. Pillement and M. Mendez Real. 'RISC-V processor enhanced with a dynamic micro-decoder unit'. In: IEEE International Conference on Electronics, Circuits and Systems. Nancy, France, 18th Nov. 2024, paper #4224. URL: https://hal.science/hal-04616772.

[54] L. Pradels, S.-I. Filip, O. Sentieys, D. Chillet and T. L. Calloch. 'FPGA-based CNN Acceleration using Pattern-Aware Pruning'. In: AICAS 2024 - IEEE 6th International Conference on AI Circuits and Systems. Abu Dhabi, United Arab Emirates: IEEE, 2024, pp. 228–232. DOI: 10.1109/AICAS59952.2024.10595865. URL: https://inria.hal.science/hal-04689673 (cit. on p. 14).

[55] L. Roquet, F. Fernandes dos Santos, P. Rech, M. Traiola, O. Sentieys and A. Kritikakou. 'Cross-Layer Reliability Evaluation and Efficient Hardening of Large Vision Transformers Models'. In: Design Automation Conference (DAC). San Francisco, United States, 23rd June 2024. URL: https://hal.science/hal-04456702 (cit. on p. 15).

[56] B. Rossigneux, V. Lorrain, I. Kucher and E. Casseau. 'Importance Resides In Activations: Fast Input-Based Nonlinearity Pruning'. In: *ICONIP 2024 Conference Proceedings*. International Conference on Neural Information Processing (ICONIP). Auckland, New Zealand, 28th Mar. 2025. URL: https://inria.hal.science/hal-04920230.

[57] M. Traiola, S. Pappalardo, A. Piri, A. Ruospo, B. Deveautour, E. Sanchez, A. Bosio, S. Saeedi, A. Carpegna and A. Savino. 'Approximate Fault-Tolerant Neural Network Systems'. In: ETS 2024 - 29th IEEE European Test Symposium. 2024 IEEE European Test Symposium (ETS). La Haye, Netherlands: IEEE, 2024, pp. 1–10. DOI: 10.1109/ETS61313.2024.10567290. URL: https://hal.science/hal-04674818 (cit. on p. 14).

[58] A. Volkova, R. Garcia, F. de Dinechin and M. Kumm. 'Hardware-optimal digital FIR filters: one ILP to rule them all and in faithfulness bind them'. In: *Proceedings of the Asilomar conference*. 2023 Asilomar Conference on Signals, Systems, and Computers. Asilomar, United States, Mar. 2024. URL: https://inria.hal.science/hal-04398268.

Conferences without proceedings

[59] W. Guilleme, Y. Helen, R. Priem, A. Kritikakou, C. Killian and D. Chillet. 'Hardening a Neural Network on FPGA through Selective Triplication and Training Optimization'. In: RADECS 2024 - RADiation and its Effects on Components and Systems Conference. Toulouse, France, 2024, pp. 1–4. URL: https://inria.hal.science/hal-04542185.

[60] M. Lemaire, D. Massicotte, J. Poupart, P. Quinton and S. Rajopadhye. 'Polyhedra at Work: Automatic Generation of VHDL Code for the Sherman-Morrison Formula'. In: Impact 2024 - 14th International Workshop on Polyhedral Compilation Techniques. Munich, Germany, 17th Jan. 2024, pp. 1–10. URL: https://inria.hal.science/hal-04401934.

[61] D. Leothaud, J.-M. Gorius, S. Rokicki, S. Rokicki and S. Derrien. 'Exploration efficace d'espace de conception pour la synthèse de haut niveau dynamique et spéculative'. In: COMPAS 2024 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Nantes, France, 2024, pp. 1–8. URL: https://hal.science/hal-04615776 (cit. on p. 11).

[62] L. Roquet, F. Fernandes dos Santos, P. Rech, M. Traiola, O. Sentieys and A. Kritikakou. 'MaxiMals: A Low-cost Hardening Technique for Large Vision Transformers'. In: RADECS 2024 - Conference is an annual on RADiation Effects on Components and Systems. Maspalomas, Canary Islands, Spain, 2024, pp. 1–5. URL: https://hal.science/hal-04736704.

[63] L. Savary, S. Rokicki and S. Derrien. 'Hardware/Software Runtime for GPSA Protection in RISC-V Embedded Cores'. In: DATE 2025. Lyon, France, 2025. URL: https://hal.science/hal-04788484.

Scientific book chapters

[64] M. Barbareschi, S. Barone, A. Bosio and M. Traiola. 'Automatic Approximation of Computer Systems Through Multi-objective Optimization'. In: *Design and Applications of Emerging Computer Systems*. Springer Nature Switzerland, 17th Aug. 2024, pp. 383–420. DOI: 10.1007/978-3-031-42478-6_15. URL: https://inria.hal.science/hal-04396685 (cit. on p. 14).

Doctoral dissertations and habilitation theses

[65] C. Ferry. 'Automating the derivation of memory allocations for acceleration of polyhedral programs'. Université de Rennes; Colorado state university, 19th Feb. 2024. URL: https://theses.hal.science/tel-04688766.

[66] C. Gernigon. 'Neural Network Quantization Methods for FPGA On-Board Pro- cessing of Satellite Images'. Université de Rennes, 18th Dec. 2024. URL: https://hal.science/tel-04883343 (cit. on p. 13).

[67] J.-M. Gorius. 'High-Level Synthesis of Instruction Set Processors'. Université de Rennes, 20th Dec. 2024. URL: https://inria.hal.science/tel-04884873.

[68] I. Krayem. 'Methods for fast exploration of manycore architectures based network-on-chip with emerging technologies'. Université de Rennes, 29th Feb. 2024. URL: https://theses.hal.science/tel-04689936.

[69] L. Pradels. 'Efficient CNN Inference Acceleration on FPGAs: A Pattern Pruning-Driven Approach'. Université de Rennes, 19th Dec. 2024. URL: https://hal.science/tel-04883465.

Reports & preprints

[70] E. Bannier, S. Castellan, S. Derrien, F. Galassi, L. Garnier, L. Hoyet, A. l'Azou, N. Lahaye, M. J.-M. Macé, O. Martineau, A. Masson, T. Maugey, B. Ninassi, E. Rohou, M. Simonin and F. Taïani. *Reducing GHG emissions from business travel: A collaborative approach at IRISA/Inria.* Groupe de travail « missions » IRISA / Centre Inria de l'Université de Rennes, Mar. 2024, pp. 1–16. URL: https://univ-rennes.hal.science/hal-04506138 (cit. on p. 36).

[71] E.-M. El Arar, M. Fasi, S.-I. Filip and M. Mikaitis. *Probabilistic error analysis of limited-precision stochastic rounding*. 2024. URL: https://hal.science/hal-04665809 (cit. on p. 12).

[72] R. Garcia. *Reproducibility Limits of Mixed-Integer Linear Programming-based methods*. 2024. URL: https://hal.science/hal-04574653.

[73] R. Garcia and A. Goldsztejn. *A computer-assisted proof that e is rational*. Mar. 2024. URL: https://hal.science/hal-04526066.

[74] P. de Oliveira Castro, E.-M. El Arar, E. Petit and D. Sohier. *Error Analysis of Sum-Product Algorithms under Stochastic Rounding*. 17th Nov. 2024. URL: https://hal.science/hal-04787542.

Other scientific publications

[75] D. Leothaud, J.-M. Gorius, S. Derrien and S. Rokicki. 'Speculative High-Level Synthesis of RISC-V Processors'. In: RISC-V Summit Europe 2024. Munich, Germany, 2024. URL: https://hal.science/hal-04615846 (cit. on p. 11).

[76] J. Pottier, M. Mendez Real, B. Le Gal and S. Pillement. 'Dynamic insertion of instructions dedicated to sidechannel attacks detection'. In: 2024 - Colloque National du GDR SoC2. Toulouse, France, 10th June 2024. URL: https://hal.science/hal-04615379.

[77] A. Rautureau, J. Paturel, M. Quinson and S. Rokicki. 'Quantifying the tiny-Small design of the SmolPhone'. In: ICT4S 2024 - 10th International Conference on ICT for Sustainability. Stockhlom, Sweden, 2024, pp. 1–4. URL: https://inria.hal.science/hal-04589322 (cit. on p. 36).

11.3 Cited publications

[78] 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 (cit. on p. 4).

[79] 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 (cit. on p. 6).

[80] 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 (cit. on p. 7).

[81] D. Deb, R. M.K. and J. Jose. 'FlitZip: Effective Packet Compression for NoC in MultiProcessor System-on-Chip'. In: *IEEE Transactions on Parallel and Distributed Systems* 33.1 (2022), pp. 117–128. DOI: 10.1109/TPDS.2021.3090315 (cit. on p. 18).

[82] 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 (cit. on p. 4).

[83] 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 (cit. on p. 4).

[84] 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 (cit. on p. 4).

[85] 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 (cit. on p. 7).

[86] 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 (cit. on p. 6).

[87] S. Mittal. 'A survey of techniques for approximate computing'. In: *ACM Computing Surveys (CSUR)* 48.4 (2016), pp. 1–33 (cit. on p. 6).

[88] A. Putnam et al. ‘A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services’. In: *ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. June 2014, pp. 13–24 (cit. on p. 6).

[89] S. Rehman et al. *Reliable Software for Unreliable Hardware: A Cross Layer Perspective*. Springer, 2016 (cit. on p. 7).

[90] N. Seifert et al. ‘Soft Error Susceptibilities of 22 Nm Tri-Gate Devices’. In: *IEEE Trans. on Nuclear Science* 59 (2012), pp. 2666–2673 (cit. on p. 7).

[91] 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 (cit. on p. 6).

[92] 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 (cit. on p. 7).