UMR IRISA

Activity Report 2025

Team TARAN

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

Joint team with Centre Inria de l'Université de Rennes

D3 – Architecture

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

Project-Team TARAN

Creation of the Project-Team: 2021 May 01

Each year, Inria research teams publish an Activity Report presenting their work and results over the reporting period. These reports follow a common structure, with some optional sections depending on the specific team. They typically begin by outlining the overall objectives and research programme, including the main research themes, goals, and methodological approaches. They also describe the application domains targeted by the team, highlighting the scientific or societal contexts in which their work is situated. The reports then present the highlights of the year, covering major scientific achievements, software developments, or teaching contributions. When relevant, they include sections on software, platforms, and open data, detailing the tools developed and how they are shared. A substantial part is dedicated to new results, where scientific contributions are described in detail, often with subsections specifying participants and associated keywords. Finally, the Activity Report addresses funding, contracts, partnerships, and collaborations at various levels, from industrial agreements to international cooperations. It also covers dissemination and teaching activities, such as participation in scientific events, outreach, and supervision. The document concludes with a presentation of scientific production, including major publications and those produced during the year.

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

A1.2.6. – 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

Contents

Project-Team TARAN **1**

1 Team members, visitors, external collaborators **5**

2 Overall objectives **7**
2.1 Context: End of CMOS . . . 7
2.2 Design Stack for Custom Hardware . . . 8
2.3 Objectives of TARAN: Facilitating Cross-Layer Optimization . . . 9

3 Research program **9**
3.1 Accelerators . . . 9
3.2 Accurate Computing . . . 10
3.3 Resilient Computing . . . 10
3.4 Embracing Emerging Technologies . . . 10

4 Application domains **11**

5 Latest software developments, platforms, open data **11**
5.1 Latest software developments . . . 11
5.1.1 Gecos . . . 11
5.1.2 SmartSense . . . 12
5.1.3 TypEx . . . 12
5.2 New platforms . . . 13
5.2.1 MPTorch: a PyTorch-based framework for simulating custom precision DNN training 13
5.2.2 Hybrid-DBT . . . 13
5.2.3 Comet . . . 13
5.2.4 MMAlpha . . . 14

6 New results **14**
6.1 High-Level Synthesis of Speculative Hardware Accelerators . . . 14
6.2 Acceleration of Error Correcting Code (ECC) . . . 15
6.3 Towards Accurate Static Power Model on Multi-Core Operating Systems . . . 15
6.4 Hardware Acceleration of Kolmogorov-Arnold Networks (KANs) . . . 15
6.5 Training Deep Neural Networks with Low-Precision Accelerators . . . 16
6.6 Compression for DNN Inference . . . 17
6.7 Design of Hardware Accelerators based on Approximate Computing (AxC) . . . 17
6.8 Reliability of ANN Hardware Accelerators . . . 18
6.9 Protection of Approximate NoC Communications Against Hardware Trojan Attacks . . . 19
6.10 Side-Channel Attacks on Embedded Artificial Intelligence . . . 20
6.11 Hardware Security . . . 20

7 Bilateral contracts and grants with industry **21**
7.1 Bilateral Contracts with Industry . . . 21
7.1.1 HLS-based hardware acceleration . . . 21
7.2 Bilateral Grants with Industry . . . 21
7.2.1 Hardware accelerators for sparse DNNs on FPGA . . . 21
7.2.2 Security of AI implementations . . . 21
7.2.3 Power consumption model for virtualization . . . 22
7.2.4 Efficient implementation of parallel applications . . . 22
7.2.5 Deep learning for video compression . . . 22
7.3 Informal Collaborations with Industry . . . 22

8 Partnerships and cooperations **22**
8.1 International initiatives . . . 22
8.1.1 Inria associate team . . . 22
8.1.2 Participation in other International Programs . . . 23
8.2 International research visitors . . . 25
8.2.1 Visits of international scientists . . . 25
8.2.2 Visits to international teams . . . 25
8.3 European initiatives . . . 25
8.3.1 Horizon Europe . . . 25
8.3.2 Other european programs/initiatives . . . 26
8.4 National initiatives . . . 27
8.4.1 ANR FASY . . . 27
8.4.2 ANR Re-Trusting . . . 27
8.4.3 ANR LOTR . . . 28
8.4.4 CYBERPROS . . . 28
8.4.5 PEPR ARSENE . . . 28
8.4.6 ANR RADYAL . . . 29
8.4.7 ANR SEC-V . . . 29
8.4.8 PEPR HOLIGRAIL . . . 30
8.4.9 PEPR ARCHI-SESAM . . . 30
8.4.10 PEPR IA - AdaptING project . . . 31
8.4.11 Inria Challenge CocoRISCo . . . 31
8.4.12 Inria Exploratory Action RobotiCore . . . 32
8.4.13 RAPID FOCH . . . 32
8.4.14 Inria Challenge OmicFinder . . . 33

9 Dissemination **33**
9.1 Promoting scientific activities . . . 33
9.1.1 Scientific events: organisation . . . 33
9.1.2 Scientific events: selection . . . 34
9.1.3 Journal . . . 34
9.1.4 Invited talks . . . 35
9.1.5 Leadership within the scientific community . . . 35
9.1.6 Scientific expertise . . . 36
9.1.7 Standardization activities . . . 36
9.2 Teaching - Supervision - Juries - Educational and pedagogical outreach . . . 36
9.2.1 Teaching administration . . . 36
9.2.2 Teaching . . . 36
9.2.3 Supervision . . . 37
9.3 Popularization . . . 38

10 Scientific production **39**
10.1 Major publications . . . 39
10.2 Publications of the year . . . 41
10.3 Cited publications . . . 45

1 Team members, visitors, external collaborators

Research Scientists

- Olivier Sentieys [Team leader, INRIA, Senior Researcher, HDR]
- Fernando Fernandes Dos Santos [INRIA, ISFP]
- Angeliki Kritikakou [INRIA, Senior Researcher, from Nov 2025, HDR]
- Nathan Leroux [CNRS, Researcher, from Oct 2025]
- Marcello Traiola [INRIA, Researcher]

Faculty Members

- Alice Chillet [UNIV RENNES, Associate Professor, from Sep 2025]
- Daniel Chillet [UNIV RENNES, Professor, HDR]
- Angeliki Kritikakou [UNIV RENNES, Associate Professor, until Nov 2025, Innovation Chair, IUF, HDR]
- Bertrand Le Gal [UNIV RENNES, Associate Professor, HDR]
- Patrice Quinton [ENS RENNES, Emeritus]
- Simon Rokicki [ENS RENNES, Associate Professor]

Post-Doctoral Fellows

- Rafael Billig Tonetto [UNIV RENNES, Post-Doctoral Fellow]
- El-Mehdi El Arar [UNIV RENNES, Post-Doctoral Fellow, until Aug 2025]
- Remi Garcia [UNIV RENNES, Post-Doctoral Fellow, until Nov 2025]

PhD Students

- Valentin Abgrall [UNIV RENNES, from Dec 2025]
- Oussama Ait Sidi Ali [SAFRAN, CIFRE, until Apr 2025]
- Hamza Amara [UNIV RENNES, until Sep 2025]
- Herinomena Andrianatrehina [INRIA]
- Habib Aouinti [INRIA, from Mar 2025]
- Gaetan Barret [ORANGE, CIFRE, until Nov 2025]
- Sami Ben Ali [INRIA, until Mar 2025]
- Noam Bires [UNIV RENNES, from Oct 2025]
- Nour Chiboub [INRIA, from Apr 2025]
- Benoit Coqueret [THALES, CIFRE, until Oct 2025]
- Taha El Idrissi [INRIA, from May 2025]
- Sohaib Errabii [INRIA]
- Romain Facq [INRIA]

- Alexandros Farmakis [INRIA, from Nov 2025]
- Thomas Feuilletin [UNIV RENNES, from Sep 2025]
- Wilfread Guillemé [INRIA, until Sep 2025]
- Gwendal Le Martin [UNIV NANTES, from Dec 2025]
- Dylan Leothaud [UNIV RENNES]
- Guillaume Lomet [INRIA]
- Elyakim Mirande-Ney [UNIV RENNES, from Dec 2025]
- Vishesh Mishra [Indian Institute of Technology Kanpur, from Jul 2025 until Nov 2025]
- Hadrien Moulherat [INRIA, from Oct 2025]
- Pegdwende Nikiema [UNIV RENNES]
- Leo Pajot [KEYSOM SAS, CIFRE]
- Lucas Roquet [UNIV RENNES]
- Baptiste Rossigneux [CEA, until Oct 2025]
- Marwa Saad [UNIV RENNES, from Sep 2025]
- Louis Savary [INRIA, until Nov 2025]
- Nesrine Sfar [UNIV RENNES]
- Mario Wagner [UNIV RENNES, from Oct 2025]
- Anis Yagoub [KEYSOM SAS, CIFRE]

Technical Staff

- Sami Ben Ali [INRIA, Engineer, from Apr 2025]
- Baptiste Bernier [INRIA, Engineer, from Jun 2025]
- Ewen Coquio [INRIA, Engineer, from May 2025]
- Jean-Michel Gorius [UNIV RENNES, Engineer, from Feb 2025 until Sep 2025]
- Wilfread Guillemé [UNIV RENNES, Engineer, from Oct 2025]
- Amélie Marotta [INRIA, Engineer, from Oct 2025]
- Pierre Nozet [INRIA, Engineer, from Jun 2025]
- Joseph Paturel [INRIA, Engineer]
- Vu Hoang Anh Pham [UNIV RENNES, Engineer, from Sep 2025]
- Leo Pradels [INRIA, Engineer, from Feb 2025]
- Maela Schmidt [INRIA, Engineer, from Jun 2025]
- Etienne Tehrani [INRIA, Engineer, until May 2025]

Interns and Apprentices

- Edouard Aubert [ENS RENNES, Intern, from Oct 2025]
- Vincent Badetti [ENS RENNES, Intern, from Mar 2025 until Aug 2025]
- Romain De Beaucorps [ENS RENNES, from Feb 2025 until May 2025]
- An-Pang Fang [UNIV RENNES, Intern, from Dec 2025]
- Thomas Feuilletin [UNIV RENNES, Intern, until Jul 2025]
- Harry Hauer [UNIV RENNES, Intern, from May 2025 until Aug 2025]
- Remi Loison [UNIV RENNES, Intern, from May 2025 until Aug 2025]
- Noe Mercier-Brosse [UNIV RENNES, Intern, from Oct 2025]
- Mathys Minsac [UNIV RENNES, Intern, from Apr 2025 until Aug 2025]
- Elyakim Mirande-Ney [UNIV RENNES, Intern, from Mar 2025 until Jul 2025]
- Hadrien Moulherat [INRIA, Intern, from Mar 2025 until Aug 2025]
- Maxwell Pirtle [ENS RENNES, until May 2025]
- Erwan Tanguy-Legac [ENS RENNES, until May 2025]
- Gauvain Thomas [INRIA, Intern, from Mar 2025 until Aug 2025]
- Emile Thuillier [ENS RENNES, Intern, from May 2025 until Jul 2025]
- Maxime Zingraff [ENS RENNES, Intern, from Oct 2025]

Administrative Assistant

- Nadia Derouault [INRIA]

Visiting Scientists

- Gayathri Ananthanarayanan [Indian Institute of Technology Dharwad (India), from May 2025 until Jun 2025]
- Omkar Shende [Indian Institute of Technology Dharwad (India), from May 2025 until Jun 2025]

2 Overall objectives

Energy efficiency has now become one of the main requirements for virtually all computing platforms [85]. 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 break down [84, 83], 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 [80]. 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: A. Chillet, D. Chillet, A. Kritikakou, B. Le Gal, N. Leroux, 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% [89], and the increasing need for acceleration of deep learning workloads [92].

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 [81, 87]. 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, 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 [19]. 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 [88]. 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 [82].

3.3 Resilient Computing

Key Investigators: 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 [86], leading to dramatically reduced yield rates [90]. Transient errors caused by particles or radiations will also more and more often occur during execution [93, 91], 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 [23, 22, 16], 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, N. Leroux, 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 [25] and we will continue to explore new application domains where HW acceleration is pertinent.

5 Latest software developments, platforms, open data

5.1 Latest software developments

5.1.1 Gecos

Name: Generic Compiler Suite

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

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

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

News of the Year: 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: Simon Rokicki

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

Partner: ENS Rennes

5.1.2 SmartSense

Name: Sensor-Aided Non-Intrusive Load Monitoring

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

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

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

Contact: Olivier Sentieys

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

5.1.3 TypEx

Name: Type Exploration Tool

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

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

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

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

Contact: Olivier Sentieys

Participant: Olivier Sentieys

5.2 New platforms

5.2.1 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

5.2.2 Hybrid-DBT

Participants: Simon Rokicki, Louis Savary.

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

5.2.3 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

5.2.4 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

6 New results

6.1 High-Level Synthesis of Speculative Hardware Accelerators

Participants: 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. One of the techniques we have developed in this domain is Speculative Loop Pipelining [8], which allows to generate speculative accelerators that are more complex than manually designed ones, and even processors.

In 2025, our work continued to push the boundaries of speculative loop pipelining, focusing on improving the SpecHLS toolchain to address these challenges. The achievements can be separated into two different categories:

- Improving the SpecHLS toolchain by optimizing the generated hardware, thus reducing the cost of the generated hardware. In a first contribution, we proposed a methodology to reduce, merge or even remove recovery logic in certain situations [47]. A second contribution enables speculative hardware sharing in the generated hardware: a single, constrained-resource can be speculatively shared by two different operations, triggering a stall whenever the two operations are used in a single iteration [48].
- Using the abstraction introduced by SpecHLS to develop new analysis. We demonstrated that it is possible to extract an abstract model of the generated CPU pipeline, and to use it to compute a safe estimation of the WCET of an application running on that pipeline [43].

6.2 Acceleration of Error Correcting Code (ECC)

Participants: Bertrand Le Gal.

Error correction codes (ECC) are essential parts of communication systems, helping to increase transmission reliability. The algorithms used to decode these codes are computationally complex, making their real-time implementation challenging, especially when energy constraints are added. The flexibility constraints of current communication systems are driving a transition from dedicated hardware architectures (ASICs) to software solutions based on the use of programmable CPU-type architectures for algorithm execution. However, the operations implemented in ECC decoding algorithms, even if they are not very complex, do not match well with the ISAs of conventional processors. To improve the performance of these algorihtms on RISC-V processors, we analyzed the ECC algorithms implemented in current communication standards (LDPC, polar codes, and turbo codes), as well as various strategies for implementing their algorithmic operations [35]. This work led to the proposal of scalar and SIMD extensions to the basic RISC-V ISA, enabling substantial gains in decoding throughput at the cost of a controlled increase in hardware complexity of the cores. This work was extended to another family of error-correcting codes (LDPC-NB), for which we initially worked on algorithmic simplifications [30].

6.3 Towards Accurate Static Power Model on Multi-Core Operating Systems

Participants: Gaëtan Barret, Daniel Chillet.

Central Processing Unit (CPU) power consumption can be divided into two components: dynamic power, which is directly related to CPU activity, and static power, which is linked to transistor characteristics. While dynamic power has been extensively studied and is commonly modelled in software using performance counters, static power has received significantly less attention. Existing static power models for real hardware generally rely on analytical formulas describing the phenomenon without providing the associated parameters used in the study, thereby preventing the research community from replicating the results.

In [39], we introduce a methodology that leverages: (1) carefully tuned benchmarks to generate consistent heat on a CPU; (2) a set of statistical tools to analyse thermal traces and perform correlation analysis with other experimental parameters; and (3) a static power model accompanied by an explicit table of associated parameters. Designed with reproducibility in mind, this study provides the community with clear and explicit steps to establish a more transparent foundation for studying static power on real-world hardware.

In [40], we propose a model of the total power consumption (dynamic and static) associated with transmitting data across the memory hierarchy, from Dynamic Random Access Memory (DRAM) to CPU caches. This study consists of: (1) using a tuned benchmark to generate precise and controlled memory activity; (2) analysing the collected data to correlate power consumption with experimental parameters; and (3) applying different families of algorithms, informed by the analysis, to model the power cost of data transmission. By combining both methodologies, we aim to provide the research community with reproducible approaches to observing and modelling specific components of system power consumption. Ultimately, this work seeks to enhance existing energy management solutions, such as Linux's internal energy model and cloud orchestration frameworks.

6.4 Hardware Acceleration of Kolmogorov-Arnold Networks (KANs)

Participants: Sohaib Errabii, Marcello Traiola, Olivier Sentieys.

Kolmogorov-Arnold Networks (KANs) have garnered significant attention for their promise of improved parameter efficiency and explainability compared to traditional Deep Neural Networks (DNNs). KANs' key

innovation lies in the use of learnable non-linear activation functions, which are parametrized as splines. Splines are expressed as a linear combination of basis functions (B-splines). B-splines prove particularly challenging to accelerate due to their recursive definition. Systolic Array (SA)-based architectures have shown great promise as DNN accelerators thanks to their energy efficiency and low latency. However, their suitability and efficiency in accelerating KANs have never been assessed. Thus, in [42], we explore the use of SA architecture to accelerate the KAN inference. We show that, while SAs can be used to accelerate part of the KAN inference, their utilization can be reduced to 30%. Hence, we propose KAN-SAs, a novel SA-based accelerator that leverages intrinsic properties of B-splines to enable efficient KAN inference. By including a non-recursive B-spline implementation and leveraging the intrinsic KAN sparsity, KAN-SAs enhances conventional SAs, enabling efficient KAN inference, in addition to conventional DNNs. KAN-SAs achieves up to 100% SA utilization and up to 50% clock cycles reduction compared to conventional SAs of equivalent area, as shown by hardware synthesis results on a 28nm FD-SOI technology. We also evaluate different configurations of the accelerator on various KAN applications, confirming the improved efficiency of KAN inference provided by KAN-SAs

6.5 Training Deep Neural Networks with Low-Precision Accelerators

Participants: Sami Ben Ali, El-Mehdi El Arar, 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 previous work, 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.

In [75], we propose a mathematically founded mixed precision accumulation strategy for the inference of neural networks. Our strategy is based on a new componentwise forward error analysis that explains the propagation of errors in the forward pass of neural networks. Specifically, our analysis shows that the error in each component of the output of a linear layer is proportional to the condition number of the inner product between the weights and the input, multiplied by the condition number of the activation function. These condition numbers can vary widely from one component to the other, thus creating a significant opportunity to introduce mixed precision: each component should be accumulated in a precision inversely proportional to the product of these condition numbers. We propose a numerical algorithm that exploits this observation: it first computes all components in low precision, uses this output to estimate the condition numbers, and recomputes in higher precision only the components associated with large condition numbers. We test our algorithm on various networks and datasets and confirm experimentally that it can significantly improve the cost–accuracy tradeoff compared with uniform precision accumulation baselines [33].

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 [41], 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.

6.6 Compression for DNN Inference

Participants: Baptiste Rossigneux, Rémi Garcia, Léo Pradels, Habib Aouinti, Olivier Sentieys.

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.

Many computer vision tasks use convolutional neural networks (CNNs). These networks have a significant computational cost and complex implementations, in particular on embedded systems. A common way to implement CNNs on integrated circuits is to use low-precision quantized weights and activations instead of de facto floating-point (FP) ones. This is important to reduce the implementation cost. However, this has a drawback regarding accuracy, and Quantization-Aware Training (QAT) is one of the most popular approaches to mitigate this issue. In [44], we introduce an multiplierlessaware training approach that significantly reduces hardware resource consumption. We propose to incrementally fix weights to their current value based on their implementation cost. To compute this cost, we base our approach on Multiple Constant Multiplication (MCM) shift-and-add solving technique. With this idea, we show a global implementation cost reduction by around 25% w. r. t. a vanilla QAT approach without hardware usage in the loop. Compared to state-of-the-art multiplierless-aware training methods, the network accuracy of our designs is closer to that of a vanilla QAT baseline.

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).

In [53], our key contribution is a new procedure to linearize CNNs, in the most cost-effective way possible. 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 bypasses the usual significant computational effort to determine removable nonlinearities, whether layer-wise or channel-wise. Additionally, we provide a comprehensive analysis of network behavior during pruning, offering insights into internal damage, recovery, and effective retraining strategies.

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

Participants: Marcello Traiola.

Approximate Computing (AxC) has emerged as an effective paradigm for improving performance, power efficiency and area by relaxing exactness constraints in error-tolerant applications such as signal processing and machine learning. Traditional AxC methodologies typically explore approximation opportunities at different abstraction levels, from gate-level to Register Transfer Level (RTL), relying on systematic design-space exploration techniques to identify suitable approximation targets. For instance, the work in [31] proposes an automated RT-level AxC exploration framework that combines bit-width and statement reduction with fault injection and assertion-based accuracy evaluation. By dynamically generating assertions from simulation traces and using their violation rates as a proxy for functional accuracy, the methodology enables a genetic algorithm to efficiently rank and cluster approximation candidates, thus supporting designers in navigating the accuracy-performance trade-off in a scalable and automated manner.

Despite these advances, most AxC approaches implicitly assume generic input workloads and do not explicitly exploit knowledge of the application's input characteristics. As highlighted in [38], this limitation can prevent further optimization opportunities, especially in signal processing applications where some inputs, such as filter coefficients, are constants or follow known distributions. The concept of Input-Aware Approximation (IAA) leverages this information to enable additional approximation dimensions, yielding significant savings in area and power. However, prior IAA techniques have largely relied on ad-hoc and non-automatic design processes, which restrict their scalability and broader adoption.

Taken together, the contributions of [38] and [31] underscore the need for AxC methodologies that are both input-aware and systematic. While [31] demonstrates how automated RTL exploration can be guided by formal assertions and fault-based accuracy metrics, [38] shows that incorporating workload-specific information can substantially enhance the effectiveness of approximation. Integrating input-aware concepts into automated RTL design exploration frameworks represents a promising direction for future AxC research, enabling scalable, generic, and workload-sensitive approximation flows that maximize efficiency gains, while maintaining controlled accuracy degradation.

6.8 Reliability of ANN Hardware Accelerators

Participants: Wilfread Guillemé, Lucas Roquet, Vishesh Vishesh, Fernando Fernandes dos Santos, Daniel Chillet, Marcello Traiola, 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.

Convolutional Neural Networks (CNN) are more often used in large applications domains, such as autonomous driving, medical systems, and aerospace. However, for these critical areas, high reliability is necessary to ensure safety and robustness. While these algorithms exhibit inherent resilience, they remain susceptible to Single-Event Effects (SEE) occurring at the hardware level. Without any protection, these events, usually induced by interactions with radiation particles, can lead to errors in electronic components, potentially causing incorrect inferences and decreasing model accuracy. Therefore, fault injection studies must be considered to define protection techniques. However, even if the CNN models are compressed through quantization and/or pruning, they remain too large for an exhaustive fault injection campaign to assess their resilience. To address these challenges, we have developed SFI4NN, a Statistical Fault Injection (SFI) framework specifically designed to evaluate the fault sensitivity of fixed-point quantized and pruned CNN architectures [45]. This framework was used to model CNN resilience as a function of the pruning rate. The obtained results enable the development of hardware hardening strategies with reduced costs that are tailored to the reliability requirements of targeted applications. Experimental results demonstrate a 96% improvement in resilience, with minimal hardware overhead compared to conventional hardening techniques such as triplication.

Recent advances in deep neural networks (DNNs) have made their reliability assessment increasingly challenging [32]. The scale of modern models has grown by orders of magnitude with the advent of large

Transformer architectures, and the input space has become significantly more diverse. Today, a single LLM may contain billions of parameters and process multimodal data such as text and images. A central question in our current research is how hardware-induced faults correlate with a model's prediction confidence. In [52], we evaluate the fault sensitivity of large Transformer-based image classifiers as a function of confidence. We show that low-confidence inputs (i.e., correct predictions with low probability) are far more likely to experience TOP-1 misclassification under faults, whereas high-confidence inputs remain substantially more robust. This trend was confirmed through neutron beam experiments and software fault simulation, with low-confidence inputs exhibiting up to a 22x increase in misclassification compared to random input selection. These results indicate that confidence is a strong predictor of vulnerability and that naïve random sampling can severely underestimate worst-case fault sensitivity. This confidence dependence also affects fault-mitigation techniques, such as value clipping. In [51], we show the limitations of activation clipping when applied to vision and language Transformers (GPT-2, BART, ViT, Swin). Although clipping effectively suppresses extreme outliers (NaN, Inf, or very large activations), many corrupted values remain within valid ranges and still propagate, leading to misclassification. Even with per-layer clipping across the entire model, misclassification rates reached 10.3%, with failures concentrated in low-confidence inputs. These results demonstrate that clipping alone cannot reliably protect modern Transformers from realistic fault effects and is least effective precisely when the model is most vulnerable.

We further explore the soft error resilience and exception avoidance in approximate floating-point computing. This gap is particularly critical in deep neural network (DNN) inference, where soft error-induced errors or exceptions can significantly affect the stability and accuracy of computations. We introduce SERA-Float [50], an approximate floating-point format resilient to soft errors. Specifically, it is designed to protect floating-point computations from soft error-induced errors and exception-triggering bit-flips. Unlike prior floating-point formats, SERA-Float protects the sign and exponent bits using error-correcting codes and relies on storing 8 valid bits of mantissa rather than performing coarse truncation. Additionally, by tracking critical bits in the floating-point representation, SERA-Float prevents overflow, underflow, and NaN exceptions. Our evaluation demonstrates that SERA-Float improves the reliability of floating-point operations during DNN inference by significantly reducing exceptions and ensuring the stability of computations. Moreover, it enables energy-efficient arithmetic by leveraging narrower arithmetic units, yielding up to 80.3% energy savings per multiplication with a 0.9% reduction in DNN inference accuracy.

6.9 Protection of Approximate NoC Communications Against Hardware Trojan Attacks

Participants: Hamza Amara, Daniel Chillet.

Network-on-Chip (NoC) has emerged as the primary interconnect solution to support execution of modern applications into multi and many core architectures. Indeed, for applications such as artificial intelligence and signal processing, the volume of exchanged data has significantly increased, resulting in higher network traffic and potential performance degradation. In this context, several approximation-based techniques have been proposed to reduce communication overhead by selectively dropping payload flits at the source core and reconstructing them at the destination core. However, this approximation capability can be exploited by a hardware Trojan to maliciously drop additional payload flits without being easily detected. Such attacks generally lead to increased network traffic, as the destination core becomes unable to reconstruct payload flits and consequently requests their retransmissions.

We address this threat by proposing a protection technique, named DyEKF, specifically designed for approximate NoC. DyEKF relies on a dynamic protection mechanism that maintains a trade-off between payload flit reconstruction and retransmission requests, while preserving application-level quality [36]. Flit reconstruction is performed using an Extended Kalman Filter (EKF), which provides more effective mitigation across various attack scenarios compared to state-of-the-art techniques such as linear interpolation employed in AMNoC. Experimental results demonstrate that DyEKF significantly mitigates the impact of hardware Trojan attacks on the CIFAR-10 database, achieving up to a 2× reduction in retransmission request rates compared to linear interpolation-based reconstruction methods, while maintaining comparable application-level quality.

6.10 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 Benoit Coqueret's Thesis, 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 on FPGAs, with deployment simplified by ready-to-use frameworks and pre-trained models. However, this convenience makes them vulnerable to malicious actors seeking to reverse engineer valuable Intellectual Property (IP) through Side-Channel Attacks (SCA). In [49], we propose a methodology for recovering the hardware configuration of dataflow accelerators generated with the FINN framework. Through unsupervised dimensionality reduction, we reduce computational overhead compared to the state of the art, enabling lightweight classifiers to recover the parallelization (#PE and SIMD) and quantization parameters. We demonstrate an attack phase requiring only 78 ms to recover the hardware parameters with an accuracy of more than 97% on an MLP victim and 1.45 s with an accuracy of more than 99% on a CNN victim, both using FINN-based accelerators and an SVM classifier on side-channel traces, even with the accelerator dataflow fully loaded. Also, it took 236 ms (MLP) and 393 ms (CNN) to fully recover these parameters with an averaging of 10 traces. This approach offers a more realistic attack scenario than existing methods. Compared to SoA attacks based on *tsfresh*, our method requires 86× (MLP) and 110× (CNN) less time for preparation and yields more consistent results across all parameters, even without averaging traces.

6.11 Hardware Security

Participants: Louis Savary, Herinomena Andrianatrehina, Bertrand Le Gal, Simon Rokicki.

The team also contributed to hardware security, working to improve the security of different types of microarchitectures: improving the control flow integrity (CFI) in embedded processors, or protecting high-performance speculative microarchitectures against Spectre-like attacks.

Embedded processors can host critical applications but are more exposed to physical attacks such as fault injection. Protecting the software being executed is a critical challenge. Many of the proposed techniques to enforce the CFI of the executed software require to recompile the application to insert new instructions or new metadata. We proposed a dynamic technique which consists in dynamically analysing the binary application just-in-time, building the code signature and verifying it during the execution. This technique offers a protection close to existing techniques, without modifying the executed binaries [55]. A second contribution improves this technique by reducing the performance overhead caused by the dynamic analysis. The main idea is to move part of the analysis at install time (i.e., ahead of time) [54].

Since the divulgation of the Spectre attack in 2018, high peformance microarchitectures based on speculative execution have been studied in depth. This resulted in many different variations of the original attack, which exploit different speculation mechanisms to trigger the information leakage, or different

side-channel to effectively leak it. In the meantime, the proposed countermeasures are either too expensive, or lack generality. Another problem with existing countermeasures is the difficulty to reproduce them, and to compare them. To adress this issue, we proposed new FENCE instruction variants allowing selective speculation on out-of-order processors [37]. These instructions are then used to implement many different policies to protect against spectre attacks. Theses policies are evaluated in a common environment, and compared fairly, which demonstrated a trade-off between security and performance.

In parallel with attacks targeting the processor and its architecture directly, various architectural elements can be attacked, such as cache memories, to extract information through side channel attacks. These attacks can be carried out simply using malicious applications running on the processor. To minimize the cost of countermeasures, e.g., in terms of their impact on application execution time and/or the energy consumed by the system, it is necessary to detect attacks so that countermeasures are only triggered when necessary. The work presented in [34] proposes integrating a micro-decoding unit into the processor pipeline that allows instruction sequences to be injected into the instruction stream on a periodic basis. This reprogrammable and flexible micro-decoding unit, which has a low hardware cost feature, has been successfully deployed to detect flush and reload and ROP attacks in real time using the processor's HPC counters.

7 Bilateral contracts and grants with industry

7.1 Bilateral Contracts with Industry

7.1.1 HLS-based hardware acceleration

Participants: Olivier Sentieys, Joseph Paturel, 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.

7.2 Bilateral Grants with Industry

7.2.1 Hardware accelerators for sparse DNNs on FPGA

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.

7.2.2 Security of AI implementations

Participants: Olivier Sentieys, Benoit Coqueret.

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

7.2.3 Power consumption model for virtualization

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.

7.2.4 Efficient implementation of parallel 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.

7.2.5 Deep learning for video compression

Participants: Taha El Drissi, Olivier Sentieys.

InterDigital is funding the PhD thesis of Taha El Drissi on Deep Learning for video compression.

7.3 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).

8 Partnerships and cooperations

8.1 International initiatives

8.1.1 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.

TCHE

Title: Transformers' Reliability for Safety-Critical Applications on Graphic Processing Units

Duration: 2025 - 2027

Coordinator: Luigi Carro (carro@inf.ufrgs.br)

Partners: Federal University of Rio Grande do Sul (Brazil)

Inria contact: Fernando Fernandes dos Santos

Summary: The rapid adoption of large Transformer models in modern AI applications introduces substantial demands on system reliability, particularly for safety-critical domains such as space exploration, avionics, autonomous vehicles, and industrial automation. In such environments, systems must operate correctly under harsh conditions, including radiation exposure, thermal stress, and aging. At the same time, the computational and memory footprint of Transformers has grown by orders of magnitude, with emerging models reaching billions or even trillions of parameters. This combination of high computational density and harsh deployment conditions significantly complicates the design of reliable embedded AI systems.

Ensuring dependable execution on embedded GPU accelerators is a key challenge. GPUs provide the parallelism required for Transformer inference, but their architectural complexity makes them susceptible to transient hardware faults. Existing fault-mitigation methods designed for larger or power-hungry platforms are not directly applicable to embedded devices due to strict energy constraints. In these settings, protection mechanisms must be lightweight, practical, and aware of the computational characteristics of the model to avoid excessive overheads. In this collaborative project, we investigate strategies that combine microarchitectural fault simulation, software-level vulnerability analysis, and energy-aware reliability techniques to improve the robustness of large Transformers running on embedded GPU accelerators. The objective is to develop fault-tolerance approaches that provide strong reliability guarantees while maintaining efficiency in resource-constrained environments.

8.1.2 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.

EdgeTrain

Participants: Olivier Sentieys, Guy Lemieux (UBC).

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

Partner Institution: University of British Columbia (UBC), Vancouver (Canada)

Summary: The research direction that our collaboration 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 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, and (iii) the production of an automated design flow for the generation of custom accelerators specialized for a given deep neural network model to train.

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
- Rutherford Appleton Laboratory (U.K), neutron beam experiments at ChipIr facility.
- Institut Laue-Langevin (FR), beam experiments at TENNIS facility.
- School of Informatics, Aristotle University of Thessaloniki (Greece), Memory management, fault tolerance
- School of Automation, Southeast University (China), Fault-tolerant task scheduling onto multi-core.
- Shantou University (China), Runtime efficient algorithms for subgraph enumeration.
- Department of Electrical and Computer Engineering, University of Naples (Italy), Digital Hardware Design Space Exploration for Approximate-Computing-based Applications
- Department of Control and Computer Engineering, Politecnico di Torino (Italy), Fault tolerance of Deep Neural Network hardware accelerators
- Department of Computer Science, University of Verona (Italy), Assertion-driven Design Exploration of Approximate Hardware

8.2 International research visitors

8.2.1 Visits of international scientists

- Gayathri Ananthanarayanan, Assistant professor at Indian Institute of Technology (IIT) Dharwad (Inde) and Omkar Shende, Phd Student at Indian Institute of Technology (IIT) Dharwad (Inde), visted TARAN on May-June 2025
- Luigi Carro, Full Professor at UFRGS, Brazil, visted TARAN on September 2025

8.2.2 Visits to international teams

- In March 2025, Fernando Fernandes dos Santos visited the Department of Computer Science of Western Paraná State University, Brazil.

8.3 European initiatives

8.3.1 Horizon Europe

EuroHPC DARE

Participants: Olivier Sentieys, Simon Rokicki.

Title: DARE: Digital Autonomy for RISC-V in Europe (Specific Grant Agreement 1)

Duration: From March 1, 2025 to February 29, 2028

Partners: 45 partners from industry and academia

Inria contact: Olivier Sentieys

Coordinator: BSC

Summary: DARE will address Europe's deficit in digital autonomy for High Performance Computing and AI, by creating truly European products for European supercomputers for research and industry. The project takes advantage of the open RISC-V ecosystem, chiplet revolution and open-source software. DARE will develop and tape-out, in advanced technology, three RISC-V-based chiplets: a vector accelerator for HPC, an AI Processing Unit inference accelerator, and an HPC-focused general-purpose processor. In DARE, TARAN, along with Codasip, will investigate advanced speculative execution techniques for branch prediction and prefetching, focusing on Performance, Power, and Area (PPA) metrics.

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

WP Leader: Angeliki Kritikakou, Inria

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)

Summary: 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.

8.3.2 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.

8.4 National initiatives

8.4.1 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 - July 2026
- P.I.: K. Kritikakou, TARAN

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

8.4.2 ANR Re-Trusting

Participants: Olivier Sentieys, Angeliki Kritikakou, Marcello Traiola.

- Program: ANR PRC
- Project acronym: Re-Trusting
- Project title: REliable hardware for TRUSTworthy artificial INtelliGence
- Duration: Oct. 2021 - July 2026
- Coordinator: INL
- Other partners: LIP6, TARAN, THALES

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

8.4.3 ANR LOTR

Participants: 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, LabSTICC

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).

8.4.4 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.

8.4.5 PEPR ARSENE

Participants: Louis Savary, Herinomena Andrianatrehina, Simon Rokicki, 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.

8.4.6 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.

8.4.7 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.

8.4.8 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.

8.4.9 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.

8.4.10 PEPR IA - AdaptING project

Participants: Angeliki Kritikakou, Marcello Traiola.

- Program: PEPR IA
- Project acronym: AdaptING
- Project title: Adaptive architectures for embedded artificial INtelliGence
- Duration: Oct 2023 – Dec 2029
- Coordinator: Alberto Bosio, ECL/INL, Lyon
- Other partners: CEA-List, IETR, Lab-STICC, Inria (MALT, TARAN), LIP6

The increasing need to distribute AI applications from the cloud to edge devices is becoming a pressing concern for addressing data privacy, bandwidth limitations, power consumption reduction, and low latency requirements, especially for real-time, mission- and safety-critical applications. Consequently, there is an ongoing effort to design custom and embedded AI hardware architectures (AI-HW) that can support energy-intensive data movement, speed of computation, and large memory resources that AI requires to achieve their full potential. However, the current AI-HW architectures are mainly based on GPU, TPU, or specialized designs, which are devoted to improving the energy/performance efficiency for a specific class of AI applications, such as Convolutional Neural Networks. Thus, they are not designed to provide the high flexibility and massive parallelism needed to support a wide range of AI algorithms, including dynamic networks, Recurrent NNs, Transformers, etc. To address these limitations, the AdaptING project proposes a new architectural paradigm called adaptive architecture, which aims to make HW adaptable to any given AI application and its constraints in terms of accuracy, energy, latency, and reliability. The adaptive architecture is designed to provide flexibility, efficiency, sustainability, and reliability for embedded AI. This approach goes beyond the current state-of-the-art HW architectures and targets the next generation of AI by investigating and designing flexible, efficient, sustainable, and reliable embedded AI on adaptive architectures.

8.4.11 Inria Challenge CocoRISCo

Participants: Simon Rokicki, 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.

8.4.12 Inria Exploratory Action RobotiCore

Participants: Marcello Traiola.

- Program: Exploratory Action
- Project acronym: RobotiCore
- Project title: Integrated Hardware Acceleration : the Brain and Heart Behind Next-Generation Robots
- Duration: Nov 2025 – Oct 2028
- Coordinator: Marcello Traiola, Taran and Marco Tognon, RAINBOW

Imagine a world where small-scale robots become indispensable companions in our daily lives — aiding in medical emergencies, ensuring personal safety, and performing tasks we hadn't even imagined. RobotiCore aims at making this vision a reality by pioneering a unified hardware accelerator that slashes power consumption and boosts efficiency for autonomous systems. Traditional robotic platforms are bulky, energy-hungry, and costly, limiting their potential. But RobotiCore wants to change the game with a groundbreaking approach: a single, low-power, compact hardware solution designed to accelerate perception, planning, and control simultaneously. This innovation opens the door to frugal, scalable, and affordable robots that can operate for far longer than current ones, unlocking unprecedented possibilities. By merging cutting-edge expertise in computer architecture and robotics, RobotiCore is paving the way for a future where energy-efficient, intelligent robots seamlessly integrate into our lives.

8.4.13 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.

8.4.14 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.

9 Dissemination

Participants: Daniel Chillet, Fernando Fernandes dos Santos, Angeliki Kritikakou, Bertrand Le Gal, Simon Rokicki, Olivier Sentieys, Marcello Traiola.

9.1 Promoting scientific activities

9.1.1 Scientific events: organisation

General chair, scientific chair

- M. Traiola was the General co-chair of IEEE IOLTS 2025.
- A. Kritikakou was the General co-chair of IEEE IOLTS 2025.

Member of the organizing committees

- D. Chillet was in the Organizing Committee of Rapido, 2025.
- A. Kritikakou, O. Sentieys and M. Traiola were members of the IEEE/ACM DATE Executive Committee, 2025.
- M. Traiola was the Review and Programme Operations Chair of IEEE/ACM DATE, 2025.
- F. Fernandes dos Santos was Publication chair of IEEE IOLTS, 2025

- A. Kritikakou was Education chair of ESWEEK, 2025.
- A. Kritikakou was Publicity chair of DDECS, 2025.
- A. Kritikakou was Student Research Forum chair of ISVLSI, 2025
- A. Kritikakou was in the Organizing Committee of BITFLIP by DGA, 2025
- M. Traiola was Publication Chair of IEEE VTS, 2025
- M. Traiola was Publication Chair of IEEE DFT, 2025
- M. Traiola was Publicity co-chair of IEEE ETS, 2025

9.1.2 Scientific events: selection

Chair of conference program committees

- O. Sentieys was co-chair of the Focus Sessions at IEEE/ACM DATE Executive Committee, 2025.
- A. Kritikakou was the Program co-chair of RTNS, 2025.
- A. Kritikakou was Special Session chair of ETS, 2025.
- M. Traiola was the Program co-chair of IEEE DDECS, 2025.
- D. Chillet was the Program co-chair of Gretsi, 2025.

Member of the conference program committees

- D. Chillet was member of the technical program committee of HiPEAC Rapido, ComPAS, DASIP, ARC, ISCIT, IEEE ICM, IEEE MCSoC.
- M. Traiola was member of the technical program committee for IEEE ICCAD, IEEE VTS, IEEE ETS, IEEE IOLTS, IEEE DFT, 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 VTS, IEEE DFT, CASES, IEEE VLSI TSA, and IEEE DATE.
- A. Kritikakou was member of the technical program committee of DAC, RTSS, CASES, ECRTS, RTAS, VLSID, RTCSA, ISVLSI, ETS, SAMOS, DS-RT, AEiC, LATS, OSPERT, AI-TREATS, COMPAS.
- O. Sentieys was a member of technical program committee of IEEE/ACM ICCAD, IEEE FPL, ACM ENSSys, DSD.
- S. Rokicki is a member of program committee of IEEE/ACM DATE.

9.1.3 Journal

Member of the editorial boards

- D. Chillet is member of the Editor Board of Journal of Real-Time Image Processing (JRTIP).
- A. Kritikakou is Handling Editor for Elsevier Microprocessors and Microsystems Journal.
- A. Kritikaou is Associate editor in Elsevier Journal of Systems Architecture.
- M. Traiola is Guest editor in IEEE Design & Test.
- M. Traiola is Guest editor in IEEE Transactions on Device and Materials Reliability.

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) and ACM/Springer journals (JSA, TECS, JSC, JESCTS).
- D. Chillet was a reviewer for Microprocessors and Microsystems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Journal of Systems Architecture, ACM Transactions on Architecture and Code Optimization, and IEEE Transactions on Very Large Scale Integration Systems.
- 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, IEEE Transactions on Pattern Analysis and Machine Intelligence, 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.)

9.1.4 Invited talks

- F. Fernandes dos Santos gave an invited talk on *Challenges in Adopting Emerging AI Accelerators For Safety-Critical and Space Applications* at IEEE NSREC Short Course 2025.
- F. Fernandes dos Santos gave an invited talk (Online) on *How Radiation Can Mess with Your DNN — and What We Can Do About It!* at I workshop on design and manufacturing of integrated circuits with open source tools, Universidad Pedagógica y Tecnológica de Colombia, 2025.
- F. Fernandes dos Santos gave an invited talk on *Hardware Fault Effects in DNN Inference and Strategies for Efficient Mitigation* at Computer Science Departament at Western Paraná State University, Brazil, 2025.
- A. Kritikakou gave a keynote talk on *Building Reliable AI Systems: Rethinking Assessment Across the Stack* during AI and Infrastructures scientific days, organized by GDR-SOC2 and GDR-RSD, France, 2025
- M. Traiola gave a keynote talk titled *"Toward Adaptive Embedded Systems: from Multi-Objective Design to Runtime Adaptation"* at the 13th Prague Embedded Systems Workshop (PESW), Czech Republic, 2025
- O. Sentieys gave an invited talk on Compressing Neural Networks for Deployment or Training at InterDigital seminar, 2025.

9.1.5 Leadership within the scientific community

- D. Chillet is a member of the French National University Council in Signal Processing and Electronics (CNU - Conseil National des Universites, 61ème section) since 2019.
- D. Chillet is member of the Board of Directors of Gretsi Association.
- A. Kritikakou is a member of the French National University Council in Computer Science (CNU - Conseil National des Universites, 27ème section) since 2022 and until 11/2025.
- A. Kritikakou is co-animator of the "High performance embedded computing" topic of GDR SoC2.
- A. Kritikakou is the president of IEEE CEDA French Chapter, founded in September 2025.

- 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.
- M. Traiola is co-chair of the "Méthodologies et Outils" topic of GDR SoC2.

9.1.6 Scientific expertise

- A. Kritikakou was a reviewer for ANR France 2030.
- B. Le Gal was a reviewer for the NWO Talent Programme (Nederland)
- M. Traiola was a reviewer for ANR-DFG joint call 2025.

9.1.7 Standardization activities

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

9.2 Teaching - Supervision - Juries - Educational and pedagogical outreach

9.2.1 Teaching administration

- A. Kritikakou is a member of the Examination Committee of Industrial Engineering Sciences and Computer Engineering (SII) Aggregation.
- B. Le Gal is associate director of studies at ENSSAT Engineering Graduate School since Nov 2024.
- S. Rokicki is responsible of the second year of the computer science department at ENS Rennes

9.2.2 Teaching

- D. Chillet: Basic processor architectures, 30h, Enssat (L3)
- D. Chillet: Multimedia processor architectures, 30h, Enssat (M2)
- D. Chillet: Advanced processor architectures, 20h, Enssat (M2)
- D. Chillet: Embedded Systems 40h (M1)
- D. Chillet: Embedded Software Development, 16h, Enssat (L3)
- D. Chillet: Low-power digital CMOS circuits, 6h, 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)
- F. Fernandes dos Santos: Frugal AI at Master 2 at ISTIC alternance (39h TPs)

9.2.3 Supervision

- PhD defended: Hamza Amara, Evaluation and Protection of Compressed Network on Chip Communications Against Hardware Trojan Attacks, Dec. 2025, E. Casseau, D. Chillet, C. Killian [63].
- PhD defended: Gaetan Barret, Predictive models for the energy cost of cloud-native applications, Dec. 2025, D. Chillet [64].
- PhD defended: Sami Ben Ali, Efficient Low-Precision Training for Deep Learning Accelerators, April 2025, O. Sentieys, S. Filip [66].
- PhD defended: Benoit Coqueret, On the Complementarity of Software and Side-Channel Attacks against Deep Learning Algorithms, CIFRE Thesis with Thales, Dec. 2025, O. Sentieys, M. Carbone (Thales), G. Zaid (Thales) [67].
- PhD defended: Wilfread Guilleme, Fault-Tolerant Hardware Architectures for Artificial Intelligence Algorithms, Dec. 2025, D. Chillet, C. Killian, A.Kritikakou [68].
- PhD defended: Baptiste Rossigneux, Sparsity in neural networks: the trade-off between operation count and complexity for embedded systems, Oct. 2025, E. Casseau, I. Kucher (CEA), V. Lorrain (CEA) [72].
- PhD defended: Léo De La Fuente, Energy optimization in near-memory computing through local instruction generation, May 2025, O. Sentieys, J.-F. Christmann (CEA) [69].
- PhD defended: Seungah Lee, Efficient Designs of On-Board Heterogeneous Embedded Systems for Space Applications, Jan. 2025, A. Kritikakou, E. Casseau, R. Salvador, O. Sentieys [70].
- PhD defended: Amélie Marotta, Effects of synchronous clock glitch effect on the security of an integrated circuit, June 2025, O. Sentieys, R. Lashermes (LHS), Rachid Dafali (DGA) [71].

- PhD in progress: Valentin Abgrall, Vulnerability analysis, fault modeling, and countermeasures towards dependable real-time computing in safety-critical drone systems, December 2025, A. Kritikakou, M. Traiola.
- PhD in progress: Noam Bires, Trustworthy AI Hardware Architecture, October 2025, A. Kritikakou, M. Traiola.
- PhD in progress: Nour Chiboub, FPGA hardware acceleration of genomic data indexing, April 2025, D. Chillet, B. Le Gal.
- PhD in progress: Taha El Idrissi, Architectures for Deep Video Compression, April 2025, O; Sentieys, O. Le Meur (InterDigital).
- PhD in progress: Alexandros Farmakis, Hardware Accelerations for Unmanned Aerial Vehicle Control Algorithms, Nov. 2025, M. Traiola, M. Tognon (Rainbow), T. Belvedere (Rainbow)

- PhD in progress: Thomas Feuilletin, An HDL for Mircroarchitecture Design, Oct. 2025, S. Rokicki, I. Puaut (PACAP), S. Derrien (Lab-STICC).
- PhD in progress: Gwendal Le Martin, Enhancing security through dynamic micro-instruction decoding management, Dec. 2025, B. Le Gal, S. Pillement (IETR).
- PhD in progress: Elyakim Mirande-Ney, Dedicated DSL for modelisation, simulation and generation of heterogeneous NoCs, Dec. 2025, D. Chillet, B. Le Gal.
- PhD in progress: Hadrien Moulherat, Automated Exploration of Microarchitectures for RISC V Processors, Oct. 2025, A. Chillet. L. Zaourar (CEA), O. Sentieys.
- PhD in progress: Marwa Saad, Increasing the operational lifespan of a Multiprocessor Circuit based on the open and free RISC V architecture, September 2025, E. Casseau, D. Chillet, B. Le Gal.
- PhD in progress: Mario Wagner, Reliability of unconventional non-Von Neuman hardware accelerators, October 2025, F. Fernandes dos Santos, M. Traiola, A. Kritikakou.
- 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: Herinomena Andrianatrehina, Ensuring confidentiality in modern Out-of-Order cores, Nov 2022, S. Rokicki, R. Lashermes.
- 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: Louis Savary, Security of DBT-based processors, Sept 2022, S. Rokicki, S. Derrien.

9.3 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.

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 was released in early 2025.

Members of Taran participate to the working group at IRISA/Inria on reducing GHG emissions from business travel.

10 Scientific production

10.1 Major publications

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

[2] S. Barone, M. Traiola, M. Barbareschi and A. Bosio. 'Multi-Objective Application-driven Approximate Design Method'. In: *IEEE Access* 9 (2021), pp. 86975–86993. DOI: 10.1109/ACCESS.2021.3087858. URL: https://hal.science/hal-03257706.

[3] 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.

[4] 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.

[5] A. Bosio, S. Germiniani, G. Pravadelli and M. Traiola. 'A genetic approach for automatic AxC design exploration at RTL based on assertion mining and fault analysis'. In: *IEEE Transactions on Emerging Topics in Computing* (19th Sept. 2025), pp. 1–15. DOI: 10.1109/TETC.2025.3609050. URL: https://inria.hal.science/hal-05333873.

[6] 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.

[7] 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.

[8] 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. 14).

[9] 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.

[10] 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.

[11] S. Errabii, O. Sentieys and M. Traiola. 'KAN-SAs: Efficient Acceleration of Kolmogorov-Arnold Networks on Systolic Arrays'. In: IEEE/ACM Design, Automation & Test in Europe Conference (DATE) 2026. Verona, Italy, 20th Apr. 2026. URL: https://inria.hal.science/hal-05372642.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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. 10).

[17] 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.

[18] 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.

[19] 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. 10).

[20] V. Mishra, M. Traiola, A. Kritikakou, O. Sentieys and U. Chatterjee. 'SERA-Float: A Soft Error Resilient Approximate Floating-Point Computing Format'. In: ICCAD 2025 - ACM/IEEE International Conference On Computer Aided Design. Munich (Allemagne), Germany: IEEE, 2025, pp. 1–9. URL: https://inria.hal.science/hal-05333255.

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

[22] 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. 10).

[23] 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. 10).

[24] 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.

[25] 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. 11).

[26] 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.

[27] 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.

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

[29] 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.

10.2 Publications of the year

International journals

[30] A. Abdallah, C. Monière, B. Le Gal and E. Boutillon. 'Adaptive Multiple-Attempts Approach to Optimize Multiple-Vote Symbol-Flipping NB-LDPC Decoder'. In: *IEEE Communications Letters* (2025), pp. 1–5. DOI: 10.1109/LCOMM.2025.3623073. URL: https://hal.science/hal-05318815. In press (cit. on p. 15).

[31] A. Bosio, S. Germiniani, G. Pravadelli and M. Traiola. 'A genetic approach for automatic AxC design exploration at RTL based on assertion mining and fault analysis'. In: *IEEE Transactions on Emerging Topics in Computing* (19th Sept. 2025), pp. 1–15. DOI: 10.1109/TETC.2025.3609050. URL: https://inria.hal.science/hal-05333873 (cit. on p. 18).

[32] 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* 72.8 (Aug. 2025), pp. 2652–2661. DOI: 10.1109/tns.2025.3528764. URL: https://hal.science/hal-04887365 (cit. on p. 18).

[33] P. de Oliveira Castro, E.-M. El Arar, E. Petit and D. Sohier. 'Error Analysis of Sum-Product Algorithms under Stochastic Rounding'. In: *SIAM Journal on Scientific Computing* 47.6 (2025), B1481–B1502. DOI: 10.1137/24M1710966. URL: https://hal.science/hal-04787542 (cit. on p. 16).

[34] J. Pottier, M. Méndez Real, B. Le Gal and S. Pillement. 'DynHaMo: Dynamic Hardware-based Monitoring dedicated to Attacks Detection'. In: *ACM Transactions on Embedded Computing Systems (TECS)* 24.5s (26th Sept. 2025), pp. 1–25. DOI: 10.1145/3762646. URL: https://hal.science/hal-05169582 (cit. on p. 21).

[35] 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 (cit. on p. 15).

International peer-reviewed conferences

[36] H. Amara, C. Killian, D. Chillet and E. Casseau. 'DyEKF: Dynamic Protection of Approximate NoC Communications Against Hardware Trojan Attacks'. In: *38th IEEE International System-on-Chip Conference*. SOCC 2025 - 38th IEEE International System-on-Chip Conference. Dubai, United Arab Emirates: IEEE, 2025, pp. 1–9. URL: https://inria.hal.science/hal-05329258 (cit. on p. 19).

[37] H. Andrianatrehina, R. Lashermes, J. Paturel, S. Rokicki and T. Rubiano. 'Exploring speculation barriers for RISC-V selective speculation'. In: ARES 2025 - 20th International Conference on Availability, Reliability and Security. Ghent, Belgium, 11th Aug. 2025, pp. 1–23. URL: https://hal.science/hal-05061555 (cit. on p. 21).

[38] M. Barbareschi, S. Barone, A. Bosio, B. Deveautour, A. Piri and M. Traiola. 'Automatic generation of input-aware approximate arithmetic circuits'. In: DDECS 2025 - IEEE 28th International Symposium on Design and Diagnostics of Electronic Circuits and Systems. Lyon, France: IEEE, 2025, pp. 139–144. DOI: 10.1109/DDECS63720.2025.11006680. URL: https://inria.hal.science/hal-05333876 (cit. on p. 18).

[39] G. Barret, D. Chillet, R. Picard and J. Penhoat. 'Towards Accurate Static Power Model on Multi-Core Operating Systems'. In: RAPIDO 2025 - 17th Workshop on Rapid Simulation and Performance Evaluation for Design Optimization. Barcelona Spain, France: ACM, 2025, pp. 1–7. DOI: 10.1145/3721848.3721849. URL: https://hal.science/hal-05273576 (cit. on p. 15).

[40] G. Barret, J. Penhoat, R. Picard and D. Chillet. 'Empirical Analysis and Estimation of the Energy Cost of Data Transfers in CPU Caches in Multi-Core Systems'. In: *2025 14th Mediterranean Conference on Embedded Computing (MECO)*. MECO 2025 - 14th Mediterranean Conference on Embedded Computing. Budva, Montenegro: IEEE (Institute of Electrical and Electronics Engineers), 2025, pp. 1–4. DOI: 10.1109/MECO66322.2025.11049192. URL: https://hal.science/hal-05273552 (cit. on p. 15).

[41] S. Ben Ali, S.-I. Filip, O. Sentieys and G. Lemieux. 'MPTorch-FPGA: a Custom Mixed-Precision Framework for FPGA-based DNN Training'. In: DATE 2025 - 28th IEEE/ACM Design, Automation and Test in Europe. Lyon, France, 2025, pp. 1–6. URL: https://hal.science/hal-04882989 (cit. on p. 16).

[42] S. Errabii, O. Sentieys and M. Traiola. 'KAN-SAs: Efficient Acceleration of Kolmogorov-Arnold Networks on Systolic Arrays'. In: DATE 2026 - IEEE/ACM Design, Automation & Test in Europe Conference. Verona, Italy, 2026. URL: https://inria.hal.science/hal-05372642 (cit. on p. 16).

[43] T. Feuilletin, D. Leothaud, S. Rokicki, S. Derrien and I. Puaut. 'Automatic Extraction of Timing Models for WCET Estimation From a High-Level Synthesis Flow'. In: DATE 2026 - Design, Automation and Test in Europe Conference. Verona, Italy, Apr. 2026. URL: https://hal.science/hal-05365718 (cit. on p. 14).

[44] R. Garcia, L. Pradels, S.-I. Filip and O. Sentieys. 'Hardware-Aware Training for Multiplierless Convolutional Neural Networks'. In: ARITH 2025 - 32nd IEEE International Symposium on Computer Arithmetic. El Paso, United States: IEEE, 2025, pp. 1–8. URL: https://hal.science/hal-04949886 (cit. on p. 17).

[45] W. Guillemé, A. Kritikakou, Y. Helen, C. Killian and D. Chillet. 'Fault Tolerance in Quantized and Pruned Convolutional Neural Networks'. In: IOLTS 2025 - IEEE 31st International Symposium on On-Line Testing and Robust System Design. Ischia, Italy: IEEE, 2025, pp. 1–7. DOI: 10.1109/IOLTS65288.2025.11117099. URL: https://inria.hal.science/hal-05313661 (cit. on p. 18).

[46] M. Jenihhin, J. Raik, A. Jutman, N. Cherezova, R. Ubar, L. Miclea, S. Enyedi, I. Stefan, O. Stan, C. Corches et al. 'European Test Symposium Teams : an Anniversary Snapshot'. In: *2025 IEEE European Test Symposium (ETS)*. ETS 2025 - 30th IEEE European Test Symposium. Tallinn, Estonia: IEEE, 1st July 2025, pp. 1–48. DOI: 10.1109/ETS63895.2025.11049652. URL: https://hal.science/hal-05173907.

[47] D. Leothaud, J.-M. Gorius, S. Rokicki and S. Derrien. 'Optimizing Recovery Logic in Speculative High-Level Synthesis'. In: DAC 2025 - 62nd Design Automation Conference. San Fransisco, United States, 2025, pp. 1–7. DOI: 10.1109/dac63849.2025.11133015. URL: https://hal.science/hal-05140504 (cit. on p. 14).

[48] D. Leothaud, S. Rokicki, S. Derrien and I. Puaut. 'Area Efficient Speculative Loop Pipelining for High-Level Synthesis'. In: DATE 2026 - Design, Automation and Test in Europe Conference. Verona, Italy, Apr. 2026. URL: https://hal.science/hal-05365717 (cit. on p. 14).

[49] G. Lomet, R. Salvador, B. Colombier, V. Grosso, O. Sentieys and C. Killian. 'Side-Channel Extraction of Dataflow AI Accelerator Hardware Parameters'. In: *2025 IEEE 30th International Symposium on On-Line Testing and Robust System Design (IOLTS)*. 2025 IEEE 31st International Symposium on On-Line Testing and Robust System Design (IOLTS). Ischia, Italy, 2025, pp. 1–7. DOI: 10.1109/IOLTS65288.2025.11117043. URL: https://inria.hal.science/hal-05120223 (cit. on p. 20).

[50] V. Mishra, M. Traiola, A. Kritikakou, O. Sentieys and U. Chatterjee. 'SERA-Float: A Soft Error Resilient Approximate Floating-Point Computing Format'. In: ICCAD 2025 - ACM/IEEE International Conference On Computer Aided Design. Munich (Allemagne), Germany: IEEE, 2025, pp. 1–9. URL: https://inria.hal.science/hal-05333255 (cit. on p. 19).

[51] L. Roquet, F. Fernandes dos Santos, L. Carro and A. Kritikakou. 'Assessing the Limitations of Activation Clipping for Fault Mitigation in Vision and Language Transformers'. In: LATS 2026 - 27th IEEE Latin American Test Symposium. Florianópolis, SC, Brazil: IEEE, 2026. URL: https://hal.science/hal-05401178 (cit. on p. 19).

[52] L. Roquet, F. Fernandes dos Santos, M. Kastriotou and A. Kritikakou. 'Strategic Input Selection For Deep Neural Networks Reliability Evaluation'. In: NSREC 2025 - IEEE Nuclear & Space Radiation Effects Conference. Nashville (Tennessee), United States: IEEE, 2025, pp. 1–5. URL: https://hal.science/hal-05006303 (cit. on p. 19).

[53] B. Rossigneux, V. Lorrain, I. Kucher and E. Casseau. 'Importance Resides In Activations: Fast Input-Based Nonlinearity Pruning'. In: *ICONIP 2024 Conference Proceedings*. ICONIP 2024 - 31st International Conference on Neural Information Processing. Auckland, New Zealand, 28th Mar. 2025, pp. 1–13. URL: https://inria.hal.science/hal-04920230 (cit. on p. 17).

[54] L. Savary, S. Rokicki and S. Derrien. 'Ahead of Time Generation for GPSA Protection in RISC-V Embedded Cores'. In: ASAP 2025 - 36th IEEE International Conference on Application-specific Systems, Architectures and Processors. Vancouver (BC), Canada: IEEE, 2025, pp. 1–7. URL: https://hal.science/hal-05100014 (cit. on p. 20).

[55] L. Savary, S. Rokicki and S. Derrien. 'Hardware/Software Runtime for GPSA Protection in RISC-V Embedded Cores'. In: DATE 2025 - Design, Automation and Test in Europe Conference. Lyon, France, 2025, pp. 1–7. URL: https://hal.science/hal-04788484 (cit. on p. 20).

[56] R. B. Tonetto, M. Traiola, F. Fernandes and A. Kritikakou. 'ENFOR-SA: End-to-end Cross-layer Transient Fault Injector for Efficient and Accurate DNN Reliability Assessment on Systolic Arrays'. In: *VTS 2026 - IEEE VLSI Test Symposium*. VTS 2026 - IEEE VLSI Test Symposium. Napa Valley, CA, United States, 2026. URL: https://hal.science/hal-05487559.

Conferences without proceedings

[57] R. Carrere, G. Ferre, B. Le Gal and P. Cais. 'Software defined radio implementation optimized for nanosatellites'. In: FAR 2025 - 3rd International Conference on Flight Vehicles, Aerothermodynamics and Re-entry. Arcachon, France, 2025, pp. 1–8. URL: https://hal.science/hal-05232579.

[58] G. Didier, A. Lucas and T. Rokicki. 'Cache Attacks in Modern/Multi-Socket x86 Systems (Work in Progress)'. In: HS3 2025 - 1st Workshop on Hardware-Supported Software Security. Toulouse, France, 26th Sept. 2025, pp. 1–10. URL: https://hal.science/hal-05249476.

[59] R. Garcia and A. Lambert. 'Bug dans plusieurs solveurs de programmation linéaire en nombres entiers'. In: ROADEF 2026 - 27ème édition du congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision. Tours, France, 2026. URL: https://hal.science/hal-05450053.

[60] W. Guilleme, A. Kritikakou, Y. Helen, C. Killian and D. Chillet. 'Tolérance aux Fautes des CNNs Quantifiés et Élagués'. In: Gretsi 2025 - XXXème Colloque Francophone de Traitement du Signal et des Images. Strasbourg, France, 25th Aug. 2025, pp. 1–4. URL: https://inria.hal.science/hal-05313756.

[61] L. Pradels, R. Garcia, S.-I. Filip, O. Sentieys and D. Chillet. 'Origami : L'art du pliage appliqué à l'accélération des CNNs sur FPGA'. In: GRETSI 2025 - Groupe de Recherche et d'Etudes de Traitement du Signal et des Images. Strasboug, France, 2025, pp. 1–4. URL: https://hal.science/hal-05392504.

[62] S. S. Sahoo, B. Deveautour, M. Traiola, C. Gu, Y. Wu, A. Japa, S. Ullah and A. Kumar. 'Special Sessions -Emerging Scope and Design Challenges for Approximate Computing: Optimizing Accuracy-PPA trade-offs and Beyond'. In: *CASES - 2025 International Conference on Compilers, Architecture, and Synthesis for Embedded Systems*. ESWEEK 2025 - Embedded Systems Week. TAIPEI, Taiwan, 13th Nov. 2025, pp. 11–20. DOI: 10.1145/3742872.375705. URL: https://hal.science/hal-05404759.

Doctoral dissertations and habilitation theses

[63] H. Amara. 'Evaluation and Protection of Compressed Network-on-Chip Communications Against Hardware Trojan Attacks'. Université de Rennes, 12th Dec. 2025. URL: https://theses.hal.science/tel-05453180 (cit. on p. 37).

[64] G. Barret. 'Energy estimation of cloud-native applications'. Université de rennes, 16th Dec. 2025. URL: https://hal.science/tel-05422961 (cit. on p. 37).

[65] S. Ben Ali. 'Entraînement à basse précision pour les accélérateurs d'apprentissage profond'. Université de Rennes, 30th Apr. 2025. URL: https://theses.hal.science/tel-05480439.

[66] S. Ben Ali. 'Low-Precision Accelerators for Efficient Deep Learning Training'. Université de Rennes, 23rd Apr. 2025. URL: https://theses.hal.science/tel-05452584 (cit. on p. 37).

[67] B. Coqueret. 'On the Complementarity of Software and Side-Channel Attacks against Deep Learning Algorithms'. Université de Rennes, 17th Dec. 2025. URL: https://theses.hal.science/tel-05452592 (cit. on p. 37).

[68] W. Guillemé. 'Fault-Tolerant Hardware Architectures for Artificial Intelligence Algorithms'. Université de Rennes, 11th Dec. 2025. URL: https://inria.hal.science/tel-05452625 (cit. on p. 37).

[69] L. D. La Fuente. 'Energy optimization in near-memory computing through local instruction generation'. Université de Rennes, 27th May 2025. URL: https://theses.hal.science/tel-05264927 (cit. on p. 37).

[70] S. LEE. 'Efficient designs of on-board heterogeneous embedded systems for space applications'. Université de Rennes, 15th Jan. 2025. URL: https://theses.hal.science/tel-05454663 (cit. on p. 37).

[71] A. Marotta. 'Effects of synchronous clock glitch on the security of integrated circuits'. Université de Rennes; RENNES, 23rd June 2025. URL: https://theses.hal.science/tel-05308528 (cit. on p. 37).

[72] B. Rossigneux. 'Sparsity in neural networks : the trade-off between operation count and complexity for embedded systems'. Université de Rennes, 2nd Oct. 2025. URL: https://theses.hal.science/tel-05425842 (cit. on p. 37).

Reports & preprints

[73] R. Boëzennec, F. Fernandes dos Santos, B. Goglin, A. Kritikakou, G. Pallez, E. Rohou, O. Sentieys and M. Traiola. *Increasing the Lifetime of HPC Machines: Issues, Implications, and Open Challenges*. 2025. URL: https://hal.science/hal-05312072.

[74] G. Didier, T. Rokicki and A. Lucas. *Flush-based Cache Attacks on Modern / Multi-Socket x86 Systems*. 19th Dec. 2025. URL: https://hal.science/hal-05424273.

[75] E.-M. El Arar, S.-I. Filip, T. Mary and E. Riccietti. *Mixed precision accumulation for neural network inference guided by componentwise forward error analysis*. 2025. URL: https://hal.science/hal-04995708 (cit. on p. 16).

[76] V. Levallois, Y. Shibuya, B. Le Gal, R. Patro, P. Peterlongo and G. E. Pibiri. *Kaminari: a resource-frugal index for approximate colored k -mer queries*. 21st May 2025. DOI: 10.1101/2025.05.16.654317. URL: https://inria.hal.science/hal-05395000.

Other scientific publications

[77] T. Feuilletin, D. Leothaud, J.-M. Gorius, S. Derrien and S. Rokicki. 'Speculative High-Level Synthesis of RISC-V Processors'. In: RISC-V summit Europe 2025. Paris, France, 2025, pp. 1–1. URL: https://hal.science/hal-05140615.

[78] D. Leothaud, J.-M. Gorius, S. Rokicki and S. Derrien. 'Optimizing Recovery Logic in Speculative High-Level Synthesis'. In: DAC 2025 - Design Automation Conference. San francisco, United States, 22nd June 2025. URL: https://hal.science/hal-05140532.

[79] L. Roquet, F. Fernandes dos Santos, M. Kastriotou and A. Kritikakou. 'Strategic Input Selection for Deep Neural Networks Reliability Evaluation'. In: NSREC 2025 - IEEE Nuclear & Space Radiation Effects Conference. Nashville (Tennessee), United States, 2025. URL: https://hal.science/hal-05273538.

10.3 Cited publications

[80] 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. 8).

[81] 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. 10).

[82] 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. 10).

[83] 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. 7).

[84] 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. 7).

[85] 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. 7).

[86] 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. 10).

[87] 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. 10).

[88] S. Mittal. 'A survey of techniques for approximate computing'. In: *ACM Computing Surveys (CSUR)* 48.4 (2016), pp. 1–33 (cit. on p. 10).

[89] 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. 10).

[90] S. Rehman et al. *Reliable Software for Unreliable Hardware: A Cross Layer Perspective*. Springer, 2016 (cit. on p. 10).

[91] 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. 10).

[92] 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. 10).

[93] 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. 10).