Offer #2024-07811

PhD Position F/M Reliability Enhancement of Unconventional AI Accelerators

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Assignment

Artificial Intelligence (AI) is increasingly indispensable across various society sectors due to its potential to transform conventional applications, from smart homes to safety-critical systems like autonomous driving and space exploration. Deep neural networks (DNNs) are state-of-the-art AI methods that outperform other approaches in language processing, image and video classification, audio and radar processing, and instance segmentation [1–3]. Notably, DNNs such as OpenAI GPT-4, Meta LLaMA2, and Mistral Mixture of Experts have captivated public interest with their high accuracy.

Due to their resource-intensive nature, DNNs require powerful dedicated hardware accelerators, such as GPUs and TPUs. However, large hardware accelerators are unsuitable for embedded safety-critical systems due to their high energy consumption. New unconventional accelerator architectures like the ones based on PIM [4] and neuromorphic computing [5] have been proposed for complex DNN deployment in critical applications where power and performance are critical requirements, offering energy-efficient alternatives to traditional GPUs and TPUs. However, their reliability, particularly against radiation-induced faults, remains to be fully assessed.

In this Ph.D., we will identify hardware and software vulnerabilities in PIM accelerators for DNNs and propose fault mitigation techniques.

Main activities

Activities

The Ph.D. student will characterize the radiation-induced impact on system reliability for different DNN model architectures and how the acceleration that PIM enables impacts the final error rate. The results will be combined with software simulation data for a detailed fault propagation analysis, aiming at deploying effective hardening solutions tailored for PIM executing DNNs.

The Ph.D. student will participate in international experiments and internships at laboratories like Rutherford Appleton Laboratory in the UK and Los Alamos National Laboratory in the USA. The student will participate in conferences and international projects and have their research published in prestigious scientific venues. This will help them develop their research skills and network with professionals in their field.

Additional information about the city and the university

Rennes is a vibrant and student-friendly city in northwestern France. The city has a thriving student culture, with plenty of bars, restaurants, cultural events, and an affordable cost of living. Additionally, *Rennes is evaluated as one of the best cities to live in Europe*[6].

Rennes is home to the University of Rennes, one of the largest universities in France. The University of Rennes has a strong focus on innovation and technology. It is home to many world-renowned research institutes, including INSA, IRISA, and INRIA Rennes. These institutes offer a wide range of Ph.D. programs in computer science, covering various topics such as artificial intelligence, machine learning, data science, and hardware and software engineering. Ph.D. students in Rennes benefit from close relationships with faculty and access to state-of-the-art facilities. The students also have the opportunity to collaborate with leading researchers worldwide.

Team's LinkedIn page: TARAN's LinkedIn

References

[1] Xiaohua Zhai et al., Scaling Vision Transformers, IEEE/CVF CVPR, 2022

[2] Chong Chen, et al., Compound fault diagnosis for industrial robots based on dual-transformer networks, Journal of Manufacturing Systems, 2023

[3] Yuxin Fang, et al., EVA-02: A Visual Representation for Neon Genesis, CVPR 2023

[4] Laguna A. F. et al., In-Memory Computing based Accelerator for Transformer Networks for Long Sequences, IEEE DATE 2021

[5] Yao M. et al., Spike-driven Transformer V2: Meta Spiking Neural Network Architecture Inspiring the Design of Next-generation Neuromorphic Chips, ICLR 2024

[6] European Commission, Quality of life in European cities, 2024, https://ec.europa.eu/regional_policy/information-sources/maps/quality-of-life_en

Skills

Required skills:

- Strong knowledge of computer architecture
- HW design: VHDL/Verilog basics, HW synthesis flow
- Basic programming knowledge (C/C++, python)
- Basics of Machine Learning
- *Experience with High-Level Synthesis (HLS) is a plus*
- *Experience in fault-tolerant architectures is a plus*
- *Knowledge of compilers and LLVM is a plus*

Languages: proficiency in written English and fluency in spoken English.

Relational skills: the candidate will work in a research team, where regular meetings will be set up. The candidate has to be able to present the progress of their work in a clear and detailed manner.

Other values appreciated: Open-mindedness, strong integration skills, and team spirit.

Most importantly, we seek highly motivated candidates.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

Remuneration

Monthly gross salary amounting to 2100 euros for the first and second years and 2200 euros for the third year

General Information

- **Theme/Domain :** Architecture, Languages and Compilation
  System & Networks (BAP E)
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2024-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2024-08-10

Contacts

- **Inria Team :** TARAN
- **PhD Supervisor :**
  Fernandes Dos Santos Fernando / fernando.fernandes-dos-santos@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different

professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Please submit online : your resume, cover letter and letters of recommendation eventually

For more information, please contact Fernando Fernandes dos santos:fernando.fernandes-dos-santos@inria.fr

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.