Offer #2025-09038

PhD Position F/M Security and Enhancement for Next-gen Trusted RAG Yields using Confidential Computing

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.

Context

This Phd thesis will be funded by the Cupseli Inria Challenge (Défi Inria) and by the ANRT

The Cupseli Inria Challenge brings together 11 Inria teams distributed over 6 Inria centers and the Hive startup company based in Cannes. Hive offers a highly original data storage architecture in which data is stored in a distributed and secure manner on the spare storage resources of participants, based on a peer-to-peer structure. This structure naturally ensures scalability, resilience and voluntary sharing of data between users. Another advantage of this architecture is its positive impact on carbon emissions by using and enhancing existing resources rather than building new datacenters, which have a very high impact.

This new challenge between Hive and Inria will focus on enabling large-scale computation on the Hive infrastructure. This objective is particularly relevant in the presence of applications that are well suited to distributed execution on interconnected computing resources, even if the network is significantly less powerful than that of a supercomputer. Although the execution of highly coupled numerical simulation codes or kernels is probably beyond the capabilities of the Hive platform, the execution of fine-tuning tasks and even the training of large models are within the scope of this challenge.

The thesis will be hosted by the WIDE team located in Rennes and will be co-supervised by membrers of the Hive startup company located in Cannes.

About WIDE

The WIDE team at the Inria center at Rennes University investigates the key fundamental theoretical and practical questions posed by modern distributed computer systems. This involves exploring the inherent tension between scalability and coordination guarantees and developing novel techniques and paradigms that are adapted to the rapid and profound changes impacting today's distributed systems, both in terms of the application domains they support and the operational constraints they must meet.

About Hive

Hive is shaping the future of cloud computing by leveraging unused computing capacity to provide a decentralized, environmentally friendly, and user-empowered alternative to traditional cloud services. By utilizing distributed peer-to-peer networks, end-to-end encryption, and blockchain technologies, Hive aims to establish a more sovereign and efficient cloud ecosystem.

Regular Meetings

The project involves regular meetings with other teams and with colleagues at Hive. Most of the meeings will be online, but some occasional trips to Paris and Cannes are foreseen. The student will also travel to international conferences to present the results of his/her work.

Assignment

This PhD Thesis plans to leverage server-side confidential computing devices to secure the processes of training and inference in decentralized machine learning. Operating on distributed data, be it for inference or for training a distributed model holds the promise for more private forms of machine learning. Instead of having a single or a few service providers that collect user data, data can remain on user devices without being collected in large data silos. In spite of this, decentralized learning does not automatically guarantee privacy. Indeed, in a distributed setting any participant becomes a potential attacker.

In this setting, the integration of confidential computing into artificial intelligence (AI) systems is becoming increasingly important [1] as organizations adopt AI for handling sensitive data in sectors like healthcare, finance, and defense. Retrieval-Augmented Generation (RAG) systems, which combine information retrieval with generative models, such as large language models (LLMs), are a popular solution for tasks like document summarization, chatbot development, and recommendation engines. However, they are often deployed in cloud environments, which introduces security risks, including data leakage, intellectual property (IP) theft, and model tampering.

The brand new Intel TDX CPU or AMD SEV-SNP features provide a promising solution by securing RAG systems through the isolation of both AI models and data within trusted execution environments (TEEs), even when deployed in multi-cloud or untrusted cloud infrastructures. Confidential computing features like TDX or SEV-SNP protect sensitive workloads by encrypting data during processing (data-in-use), while also ensuring model privacy and integrity using cryptographic attestation [2,3]

The student will leverage the Wide team's expertise in Operating Systems and Hypervisors, on trusted execution environments [4], and decentralized machine learning [5,6] to design a Trusted Virtual Machine Monitor (VMM) that will manage and secure Intel TDX-enabled Retrieval-Augmented Generation (RAG) systems, addressing key research objectives. Designing a Trusted VMM presents significant challenges due to the stringent requirements for security, performance, and scalability.

A first challenge lies in ensuring that the VMM operates transparently with the guest operating systems (OS). This entails providing essential security services—such as isolation and resource management—without modifying the guest OS or requiring it to be aware of the underlying TDX-aware VMM. Ensuring this transparency is critical for compatibility with existing systems and workloads, as modifying the guest OS would introduce substantial overhead and deployment complexity.

A second challenge results from the fact that not all available hardware will be equipped with confidential-computing devices. This results from the presence of non-trusted or incompatible server-side devices. To this end, we plan to leverage clustering approaches that can make it possible to leverage groups of devices that can communicate with each other. However, naive clustering would make the system very vulnerable to attacks. As a result, we plan to leverage a hybrid solution in which the data managed by a model is split between a protected and an unprotected part, leading to an inference or training process that is split between trusted and non trusted devices.

Main activities

The student will work on the design of a Trusted Virtual Machine Monitor (VMM).

To this end, we envision the following high-level work plan.

M0-M3: The PhD student will perform a thorough state-of-the-art of confidential computing, virtual machines and machine-learning workloads.

M4-M12: The student will work on developing novel algorithms and techniques to address the first challenged outlined above, namely, ensuring transparency for the operation of the VMM with respect to the guest operating system. This will involve developing techniques for isolation and resource management that

leverage confidential-computing hardware.

M13-M24: They will then explore hybrid approaches that make it possible to combine confidential-computing devices with non-trusted or incompatible ones.

M25-M30: The student will work on combining the results of the first two projects into a scalable and reusable Virtual Machine Monitor and will test its application to machine-learning workloads such as Retrieval-Augmented Generation (RAG) systems.

M31-M36: The final months will be devoted to writing the manuscript and on finalizing the publications of the thesis results.

Skills

Excellent programming skills and a willingness to learn about new techniques (confidential computing, machine learning, low-level programming) are also crucial, as well as good writing skills and the ability to propose, present, and discuss new ideas in a collaborative setting.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment)
- Social, cultural and sports events and activities
- Access to vocational training

Remuneration

2 200€ per month

General Information

- **Theme/Domain :** Distributed Systems and middleware
- **Town/city :** Rennes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2025-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-07-31

Contacts

- **Inria Team :** WIDE

- **PhD Supervisor :**
  Frey Davide / davide.frey@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.

The keys to success

The candidate recruited for this Ph.D. should have a Master's Degree in Computer Science or equivalent, with a solid background in distributed and operating systems. Knowledge and expertise in confidential-computing hardware, and/or machine learning is a plus.

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 your CV, cover letter, and any recommandations online

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.