Offer #2025-08853

PhD Position F/M Privacy-Enabled AI Job Execution on Heterogeneous Consumer Hardware Architectures

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Level of experience : Recently graduated

Context

The position is part of a collaboration between the Hive company and the STACK team. The successful candidate will be part of the STACK team based at IMT Atlantique, Nantes.

About STACK: STACK is a research group focusing on challenges related to the management and advanced usages of Utility Computing infrastructures (i.e. Cloud, Fog, Edge, and beyond). More specifically, the team is interested in delivering appropriate system abstractions to operate massively geo-distributed ICT infrastructures, from the lowest (system) levels to the highest (application development) ones, and addressing crosscutting dimensions such as energy or security. Those infrastructures are critical for the emergence of new kind of applications related to the digitalization of the industry and the public sector (aka Industrial and Tactile Internet).

About the Hive company: 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.

Assignment

Context

Taking security into account in distributed AI work is essential to respect user privacy, be it for inference or model training [3]. This is particularly the case when using consumer computers for computing tasks, a technique known as Desktop Grid Computing [1]. This otherwise idle hardware enables low-cost execution of computing tasks. However, in the AI case, it brings privacy issues as personal training or inference data is co-located with other users' processes. One way to overcome this challenge is to employ the hardware security capabilities provided by modern processors, coprocessors, and chipsets [6,2]. The security features present in modern consumer computers include, for example, the Trusted Platform Module (TPM) chipsets integrated in most x86 motherboards, or Microsoft Pluton [5] in modern x86 processors. These modules provide numerous functions to help secure applications, such as key storage, encryption, and decryption. In addition recent processors have capabilities to securely run code in isolation, using for instance Intel SGX, SEV or TrustZone technologies [7]. TrustZone has been used successfully on AI projects in recent work [2, 4], as ARM platforms are becoming increasingly widespread in consumer computers and data centers. However, not all these security capabilities are identical, nor do they provide the same security guarantees depending on the family and model of processors and chipsets employed. Added to this is the complexity of using this type of architecture, which often requires specific programming models resulting in heavy adaptations of the executor code [4], or in other cases a very reduced performance [3].

Goals

The main objective of this PhD is to provide middleware-like software support able to distribute and run AI workloads on resources hosted on end-user computers, according to their software and hardware capabilities, while taking into account users' privacy requirements. Several research questions will be considered:

• How to ensure the privacy of AI processing on computers with heterogeneous capabilities, taking advantage of end-user hardware security capabilities?
• How to adapt AI processing to take advantage of these security capabilities?
• What trade-offs between security and performance need to be addressed?
• How to dynamically distribute jobs across participating nodes to balance those trade-offs?

The development of this middleware is based on the following steps: First, an analysis of the capabilities and limitations of TPM and TrustZone will be carried out to understand how these technologies can secure AI processing. Then, the middleware will be designed to automatically detect the hardware capabilities of computers and adjust the security level according to the available configuration. To achieve this goal, it will integrate mechanisms to isolate AI processing with TrustZone and secure the storage of sensitive data with TPM. Finally, we will develop an AI job orchestrator based on optimization models and/or heuristics to determine the most efficient deployment according to security, performance, and

machine availability needs. Experiments on devices equipped with these technologies will validate the middleware. The tests will measure the impact of dynamic adaptation on security and performance, providing data on the feasibility and effectiveness of the middleware in real-world scenarios.

Bibliography

[1] Evgeny Ivashko, Ilya Chernov, and Natalia Nikitina. “A Survey of Desktop Grid Scheduling”. In: IEEE Transactions on Parallel and Distributed Systems 29.12 (Dec. 2018), pp.2882–2895. issn: 1558-2183. doi: 10.1109/TPDS.2018.2850004. url: https://ieeexplore.ieee.org/document/8395065/ (visited on 04/29/2025).
[2] Aghiles Ait Messaoud et al. “Shielding federated learning systems against inference attacks with ARM TrustZone”. In: Proceedings of the 23rd ACM/IFIP International Middleware Conference. Middleware ’22. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 335–348. isbn: 978-1-4503-9340-9. doi: 10.1145/3528535.3565255. url: https://dl.acm.org/doi/10.1145/3528535.3565255 (visited on 11/05/2024).
[3] Fan Mo, Zahra Tarkhani, and Hamed Haddadi. “Machine Learning with Confidential Computing: A Systematization of Knowledge”. In: ACM Comput. Surv. 56.11 (June 2024), 281:1–281:40. issn: 0360-0300. doi: 10.1145/3670007. url: https://dl.acm.org/doi/10.1145/3670007 (visited on 10/14/2024).
[4] Arttu Paju et al. “SoK: A Systematic Review of TEE Usage for Developing Trusted Applications”. In: Proceedings of the 18th International Conference on Availability, Reliability and Security. ARES ’23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 1–15. isbn: 9798400707728. doi: 10.1145/3600160.3600169. url: https://dl.acm.org/doi/10.1145/3600160.3600169 (visited on 10/17/2024).
[5] Vinay Pamnani. Microsoft Pluton security processor. en-us. July 2024. url: https://learn.microsoft.com/en-us/windows/security/hardware-security/pluton/microsoft-pluton-security-processor (visited on 11/05/2024).
[6] Lianying Zhao et al. “A Survey of Hardware Improvements to Secure Program Execution”. en. In: ACM Computing Surveys 56.12 (Dec. 2024), pp. 1–37. issn: 0360-0300, 1557-7341. doi: 10.1145/3672392. url: https://dl.acm.org/doi/10.1145/3672392 (visited on 11/04/2024).
[7] Qinyu Zhu et al. “Investigating TrustZone: A Comprehensive Analysis”. en. In: Security and Communication Networks 2023.1 (2023), p. 7369634. issn:1939-0122. doi: 10.1155/2023/7369634. url: https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/7369634 (visited on 11/03/2024).

Main activities

The PhD student will conduct original researches on the topic described above, and will collaborate with colleagues in the STACK team and Hive partners.

Activities includes, but are not limited to: bibliographical synthesis, research, proof writing, software implementation, presentation of results at conferences, attending research schools, etc

Skills

Technical skills and level required :

- Solid understanding of cybersecurity principles such as threat modeling, cryptographic primitives and key management (PKI, symmetric/asymmetric encryption, hashing) and their integration in applications
- Proficiency in distributed systems
- Proficiency in systems programming (C/C++, Rust) and high level language (Python).
- Experience with AI frameworks (PyTorch, Tensorflow, ...) and software performance measurement is a plus

Languages :

- Good communication skills in English (French is a plus)

Relational skills :

- Ability to work collaboratively in an academic–industry 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) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

Remuneration

Salary gross : 2200€

General Information

- **Theme/Domain :** Distributed Systems and middleware System & Networks (BAP E)
- **Town/city :** Nantes
- **Inria Center :** Centre Inria de l'Université de Rennes
- **Starting date :** 2025-09-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-07-05

Contacts

- **Inria Team :** STACK
- **PhD Supervisor :**
  Rosinosky Guillaume / guillaume.rosinosky@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

We are looking for a candidate with a strong background in computer science and cybersecurity. The candidate must have interest in research, a high level of curiosity and tenacity. Experience or interests in AI and distributed systems are 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 online : your resume, cover letter and letters of recommendation eventually

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.