A propos du centre ou de la direction fonctionnelle

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.

Contexte et atouts du poste

Financing Project

This PhD will be done in the context of the ENGAGE DFKI-Inria (2021-2024), focused on building next-generating computing environments for artificial intelligence.


More on DFKI: [https://www.dfki.de/en/web/](https://www.dfki.de/en/web/)

Mission confiée

Introduction

As Artificial Intelligence has recently gained an unprecedented momentum in a rapidly increasing number of application areas, Deep Neural Networks (DNNs) are becoming a pervasive tool across a large range of domains, ranging from autonomous driving vehicle, industrial automation, and pharmaceutical research to name just a few.

As these neural network architectures and their training data are getting more and more complex, so are the infrastructures that are needed to execute them sufficiently fast. Hyperparameter setting and tuning, training, inference, dataset handling are operations that are all putting a growing pressure on the underlying compute infrastructure and call for novel approaches at all levels of the workflow, including the algorithmic, the middleware and deployment level, and the resource optimization level.

This thesis is proposed as part of a collaborative project established between Inria and DFKI (the German Research Center for Artificial Intelligence). The goal of this project is to leverage efficient collaboration of experts in the AI and HPC areas to address the following specific research questions:

1. How can we deal with situations where training or validation data is not available in sufficient quantity or quality, which is the case (1) if the generation of data is expensive because of sample creation or measurement costs, (2) if the cost of manual data annotation constitutes an insurmountable effort, (3) if the natural occurring data distribution is unfavorable, i.e. highly relevant situations occur only rarely, or (4) if a phenomenon has been predicted in theory, but not yet observed. Our answer here will include the concepts of parametric models and simulations (also known as Digital Reality).
2. How can the various possible deployment options of complex AI workflows on the available underlying infrastructure impact performance metrics? How can this infrastructure be best leveraged in practice, potentially through seamless integration of supercomputers, clouds, and fog/edge systems?

This thesis is focusing on the second question

Thesis proposal

The thesis will focus on the middleware and the deployment level. Our objective is to investigate various deployment strategies for complex AI workflows (e.g., potentially combining online training, simulations and inference, all in parallel and in real-time) on hybrid execution infrastructures (e.g., combining supercomputers and cloud/fog/edge systems). This requires scalable and reliable experimentation tools. To this purpose, an important objective is to propose methodologies and supporting tools enabling researchers to:

- describe in a representative way the application behavior;
- reproduce it in a reliable, controlled environment for extensive experiments; and
- understand how the end-to-end performance of applications is correlated to various algorithm-dependent or infrastructure-dependent factors.

Principales activités

The main expected outcomes are: (1) publications describing an experimental, reproducibility-oriented methodology, its validation in practice through novel insights it can enable, potentially leading to novel algorithms for parallel/continual learning and inference across the computing continuum; (2) associated underlying algorithms and 3) an adequate software framework for experiment deployment, monitoring and execution at scale on various relevant scalable infrastructures (e.g., on experimental platforms such as Grid5000 in a first stage and on hybrid infrastructures including pre-exascale HPC platforms in a second stage).

Informations générales

- **Thème/Domaine:** Calcul distribué et à haute performance
- **Ville:** Rennes
- **Centre Inria:** CRI Rennes - Bretagne Atlantique
- **Date de prise de fonction souhaitée:** 2021-10-01
- **Durée de contrat:** 3 ans
- **Date limite pour postuler:** 2021-09-30

Contacts

- **Équipe Inria:** KERDATA
- **Directeur de thèse :** Gabriel.antoniu@inria.fr

A propos d’Inria

Inria est l’institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 200 équipes-projets agiles, en général communiques avec des partenaires académiques, impliquent plus de 3500 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 180 start-up. L’institut s’efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l’économie.

L’essentiel pour réussir

- An excellent Master degree in computer science or equivalent (e.g. engineering)
- Very good communication skills in oral and written English.
- Open-mindedness, strong integration skills and team spirit

Consignes pour postuler

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

For more information, please contact alexandru.costan@inria.fr or gabriel.antoniu@inria.fr

In parallel to the online submission on the Inria web site, please send an email with a cover letter, CV, contact address of at least two references (work collaborator, internship advisor, teacher in a related field, …) and copies of degree certificates to Dr. Gabriel Antoniu (gabriel.antoniu@inria.fr) and Dr. Alexandru Costan (alexandru.costan@inria.fr).

Sécurité défense

Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :
To address these challenges, the thesis will leverage the E2Clab approach [E2Clab2020, Ros2020] initiated in the KerData team at Inria to address the needs of experimentation of workloads involving online parallel learning and inference. In addition to potential parallelization strategies for learning and inference tasks, our goal is to enable reproducible experimentation of complex AI workloads across hybrid infrastructures and help optimize deployment strategies depending on multiple factors including the application characteristics, the target performance metrics and the features of the available execution hardware. The goal is to answer questions like: How can the various possible deployment options of complex AI workflows on the available underlying infrastructure impact performance metrics? How can this infrastructure be best leveraged in practice, potentially through seamless integration of supercomputers, clouds, and fog/edge systems?

International visibility and mobility

The thesis will be conducted in strong collaboration with several partners including DFKI (contact: René Schubotz), where a pair PhD position will be provided, and Argonne National Lab, USA (contact: Bogdan Nicola). The thesis may include long research stays (1-3 months) at the partner's teams, for joint collaborative work.

How to apply?

In parallel to the online submission on the Inria web site, please send an email with a cover letter (CV, contact address of at least two references (internship, teacher in a related field, ... ) and copies of degree certificates to Dr. Gabriel Antoniu (gabriel.antoniu@inria.fr) and Dr. Alexandru Costan (alexandru.costan@inria.fr). Incomplete applications will not be considered nor answered.

References


[E2Clab2020] The E2Clab project: https://team.inria.fr/kerdata/e2clab/.


Compétences

- Advanced knowledge of computer networks, machine learning and distributed systems
- Ability and motivation to conduct high-quality research, including publishing the results in relevant venues
- Strong programming skills (e.g. C/C++, Java, Python)
- Working experience in the areas of machine learning, HPC, distributed systems is an advantage

Avantages

- Subsidised catering service
- Partially-reimbursed public transport

Rémunération

monthly gross salary amounting to 1982 euros for the first and second years and 2085 euros for the third year.