**2020-03183 - PhD Position F/M Reproducible deployment and scheduling strategies for AI workloads on the Digital Continuum**

**Type de contrat :** CDD  
**Niveau de diplôme exigé :** Bac + 5 ou équivalent  
**Fonction :** Doctorant  
**Niveau d'expérience souhaité :** Jeune diplômé

**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 institutes, etc.

**Contexte et atouts du poste**

**Financing Project**

This PhD will be done in the context of the ACROSS EuroHPC project (2021-2023), focused on enabling efficient execution of complex workflows combining simulation, analytics and learning across hybrid infrastructures (HPC/cloud/edge).

**Mission confiée**

**Introduction**

As Artificial Intelligence has recently gained an unprecedented momentum in a rapidly increasing number of application areas, Deep Neural Networks (DNN) are becoming a pervasive tool across a large range of domains, including 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 level, the middleware and deployment level, and the resource optimization level.

**Thesis proposal**

In this thesis we focus on the middleware and the deployment level. Understanding end-to-end performance of complex AI workloads deployed on a digital continuum that may include hybrid resources (HPC/systems, clouds, edge devices) is challenging. This breaks down to conciliating many, typically contradicting constraints with low-level infrastructure design choices. One important challenge is to enable accurate, reproducible experimental investigation of relevant behaviors of a given application workflow and representative settings of the physical infrastructure. This includes automated experiment configuration at scale based on a scenario deployments previously identified, experiment execution on large testbeds (e.g., Grid5000), metrics collection and analysis, management of experimental artifacts to ensure repeatability, reproducibility and reproductibility.

**Principales activités**

To address these challenges, we will define an experimental framework and a methodology leveraging the EZClab approach [EZClab2020, Ros2020] initiated in the KerData team at Inria, and extend it to cover the complete computing continuum. In particular, EZClab will be extended GPU virtualization, containerization or the support for microservice architectures. 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?

The main expected outcomes are: (1) an experimental, reproducibility-oriented methodology and its validation in practice through novel insights it can enable (e.g., through the experimentation of alternative scheduling strategies), and (2) an associated underlying software framework for experiment deployment, monitoring, and execution at scale on various relevant scalable infrastructures.

**International visibility and mobility**

The thesis will be conducted in collaboration with several partners including DFRI (René Schubotz) and the University of Düsseldorf (Michael Schöttner).

**References**


**Informations générales**

- **Thème/Domaine :** Calcul distribué et à haute performance  
- **Ville :** Rennes  
- **Centre Inria :** CRI Rennes - Bretagne Atlantique  
- **Date de prise de fonction souhaitée :** 2021-01-01  
- **Durée de contrat :** 3 ans  
- **Date limite pour postuler :** 2021-03-03

**Contacts**

- **Equipe Inria :** KERDATA  
- **Directeur de thèse :** Costan Alexandru / Costan.Alexandru@inria.fr

**A propos d'Inria**

Inria is the national institute for research dedicated to the sciences of information. It employs 2600 personnel, 750 PhDs and students, and 200 international projects. Inria's teams are involved in a large number of domains, including 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 level, the middleware and deployment level, and the resource optimization level.

**Mission confiée**

**Introduction**

As Artificial Intelligence has recently gained an unprecedented momentum in a rapidly increasing number of application areas, Deep Neural Networks (DNN) are becoming a pervasive tool across a large range of domains, including 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 level, the middleware and deployment level, and the resource optimization level.

**Thesis proposal**

In this thesis we focus on the middleware and the deployment level. Understanding end-to-end performance of complex AI workloads deployed on a digital continuum that may include hybrid resources (HPC/systems, clouds, edge devices) is challenging. This breaks down to conciliating many, typically contradicting constraints with low-level infrastructure design choices. One important challenge is to enable accurate, reproducible experimental investigation of relevant behaviors of a given application workflow and representative settings of the physical infrastructure. This includes automated experiment configuration at scale based on a scenario deployments previously identified, experiment execution on large testbeds (e.g., Grid5000), metrics collection and analysis, management of experimental artifacts to ensure repeatability, reproducibility and reproductibility.

**Principales activités**

To address these challenges, we will define an experimental framework and a methodology leveraging the EZClab approach [EZClab2020, Ros2020] initiated in the KerData team at Inria, and extend it to cover the complete computing continuum. In particular, EZClab will be extended GPU virtualization, containerization or the support for microservice architectures. 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?

The main expected outcomes are: (1) an experimental, reproducibility-oriented methodology and its validation in practice through novel insights it can enable (e.g., through the experimentation of alternative scheduling strategies), and (2) an associated underlying software framework for experiment deployment, monitoring, and execution at scale on various relevant scalable infrastructures.

**International visibility and mobility**

The thesis will be conducted in collaboration with several partners including DFRI (René Schubotz) and the University of Düsseldorf (Michael Schöttner).

**References**

Compétences
- Strong knowledge of computer networks and distributed systems
- Knowledge on storage and (distributed) file 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 Big Data management, Cloud computing, HPC, 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