Offer #2025-08839

PhD Position F/M Large-Scale Artificial Intelligence Inference Optimization in Distributed Cloud Environments

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

Level of experience : Recently graduated

About the research centre or Inria department

The Inria center at the University of Bordeaux is one of the nine Inria centers in France and has about twenty research teams.. The Inria centre 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 SMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute...

Context

This thesis subject is proposed in the context of a collaboration between Inria and hive (https://www.hivenet.com). hive provides Hivenet, a fully distributed cloud infrastructure based on the devices of community members. In other words, members of Hivenet share their own spare computing and storage to the cloud, which then benefits other users. An advantage of this architecture is the positive impact on carbon emissions by using existing resources instead of building new data centers.

Our focus lies on the emergence of new applications that naturally fit well with distributed execution, such as the inference of large-scale artificial intelligence (AI) models that require a large number of independent tasks to be executed. These applications are well suited to distributed execution on interconnected computing resources, even if the network is significantly less powerful than that of a supercomputer.

In the context of large-scale AI inference, several issues need to be addressed:

- Each token generation must a priori pass through several graphics processing units (GPUs), each storing different parts of the model. There is a problem of model partitioning [1][2] and also a problem associated with the construction of inference paths to minimize latency (making groups of nearby resources);
- The amount of inference required will naturally vary over time, and the set of resources made available to the computation will also vary. There is therefore a static planning problem (deciding which resources are likely to participate and storing the models there) and a dynamic problem (how to allocate new requests);
- Depending on the models used, some of the inference tasks are naturally placed on certain resources (for example, because previous tokens have been generated there). In terms of fault tolerance, the computation can easily be restarted (knowing which tokens have been generated), but at a high cost. This raises problems of resource allocation depending on resource availability statistics.

Assignment

The aim of this PhD thesis is to propose and adapt techniques to improve the latency, throughput, or resource utilization of large-scale AI inference in a fully distributed cloud environment.

The work will focus on complex models (such as LLMs) that require multiple GPUs due to their memory requirements. These GPUs will also provide a complex computing platform, given their potential heterogeneity, volatility, and geographical distribution in the cloud infrastructure.

The optimization techniques to be proposed will be concentrated in scheduling and resource allocation methods to optimize the execution time of the inference tasks and their communication.

An objective in this collaboration with hive would be to deploy the optimized large-scale models in:

- compute nodes from the same supplier (with homogeneous desktop-grade GPUs available close together); and
- compute nodes from different community members (where the communication latency between suppliers has to be taken into account, and GPU models are more heterogeneous).

In the context of this topic, we also consider it relevant to mention its similarities to data-stream processing (DSP). More precisely, DSPs are models representing the continuous processing of data (akin to token generation), which passes through various operators (replicated as required) represented by a DAG. There is a large literature on the placement of operators in such models, with optimization of various metrics [3] (latency, throughput, communications, scaling, fault tolerance, etc.), although most of it concerns the case of execution in clouds [4] (fewer resource constraints and less heterogeneity).

Main activities

- Bibliographical research (scheduling and AI literature)
- Modeling of AI inference (for its optimization)
- Design, programming and validation of algorithms
- Integration of software components or modules into existing software frameworks
- Publication of research results in scientific papers
- Presentation and dissemination of the results at conferences and workshops

Additional activities:

- Attendance at the mandatory Doctoral School courses (EDMI - Univ. Bordeaux)
- Optional participation in Master-level internship student advising
- Optional teaching a the University of Bordeaux or at the ENSEIRB-MATMECA Engineering School

Skills

- Intermediary knowledge of high-performance computing and Cloud computing;
- Beginner knowledge of algorithms and optimization problems (higher proficiency is a bonus);
- Good knowledge of LLMs will be appreciated;
- Good level of software development under UNIX-like operating systems;
- Good writing skills;
- Willingness to work in a diverse and international environment.

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 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

The gross monthly salary will be 2200€, then 2300€ from 2026 (before social security contributions and monthly witholding tax).

General Information

- **Theme/Domain :** Distributed and High Performance Computing System & Networks (BAP E)
- **Town/city :** Talence
- **Inria Center :** Centre Inria de l'université de Bordeaux
- **Starting date :** 2025-10-01
- **Duration of contract :** 3 years
- **Deadline to apply :** 2025-06-30

Contacts

- **Inria Team :** TOPAL
- **PhD Supervisor :**
  Lima Pilla Laercio / laercio.lima@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

- A liking for algorithm design and optimization problems;
- Curiosity and appetence for exploration and experimentation;
- Openness to work in a team;
- Motivation to implement ethical and rigorous scientific practices.

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

If you are interested by this job, please could you apply on website jobs.inria with the following documents :

- cv
- cover letter

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.