Offre n°2020-02443

Doctorant F/H Deep Learning Beyond the Edge: Chasing Mobile Streams

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Autre diplôme apprécié : Master of Science or Engineering

Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle

Le centre Inria Rennes - Bretagne Atlantique est un des huit centres d'Inria et compte plus d'une trentaine d'équipes de recherche. Le centre Inria est un acteur majeur et reconnu dans le domaine des sciences numériques. Il est au cœur d'un riche écosystème de R&D et d'innovation : PME fortement innovantes, grands groupes industriels, pôles de compétitivité, acteurs de la recherche et de l'enseignement supérieur, laboratoires d'excellence, institut de recherche technologique.

Contexte et atouts du poste

- **Advisors: Alexandru Costan, Gabriel Antoniu (KerData team) and Bogdan Nicolae (Argonne National Laboratory, USA)**
- **Main contacts: alexandru.costan(at) inria.fr, gabriel.antoniu (at) inria.fr, bogdan.nicolae (at) acm.org**
- **Collaboration context: JLESC International Laboratory on Extreme-Scale Computing**
- **Expected start date: October 1st, 2020**
- **Application deadline: as early as possible, no later than March 12, 2020**

Location and Mobility

The thesis will be mainly hosted by the KerData team at Inria Rennes Bretagne Atlantique. It will include collaborations with **Argonne National Lab, USA** (which provides one of the target applications, where the student is expected to be hosted for a 3-month internship). Rennes is the capital city of Britanny, in the western part of France. It is easy to reach thanks to the high-speed train line to Paris. Rennes is a dynamic, lively city and a major center for higher education and research: 25% of its population are students.

The KerData team in a nutshell for candidates

- As a PhD student mainly hosted in the KerData team, you will join a dynamic and enthusiastic group, committed to top-level research in the areas of High-Perfomance Computing and Big Data Analytics. Check the team's web site: https://team.inria.fr/kerdata/.
- The team is leading multiple projects in top-level national and international collaborative environments, e.g., the JLESC international Laboratory on Extreme-Scale Computing: https://jlesc.github.io. It has active collaborations with top-level academic institutions all around the world (including the USA, Mexico, Spain, Germany, Japan, Romania, etc.). The team has close connections with the industry (e.g., Microsoft, Huawei, Total).
- The KerData team's publication policy targets the best-level international journals and conferences of its scientific area.The team also strongly favors experimental research, validated by implementation and experimentation of software prototypes with real-world applications on real-world platforms, e.g., clouds such as Microsoft Azure and some of the most powerful supercomputers in the world.

Why joining the KerData team is an opportunity for you

- The team's top-level collaborations strongly favor successful PhD theses dedicated to solving challenging problems at the edge of knowledge, in close interaction with top-level experts from both academia and industry.

  - To follow the career of our former PhD students, have a look here: https://team.inria.fr/kerdata/team-members/.

- **The KerData team is committed to personalized advising and coaching** , to help PhD candidates train and grow in all directions that are critical in the process of becoming successful, top-level researchers.

- You will have the opportunity to present your work in top level venues where you will meet the best experts in the field.

- **What you will learn.** Beyond learning how to perform meaningful and impactful research, you will acquire useful skills for communication both in written form (how to write a good paper, how to design a convincing poster) and in oral form (how to present their work in a clear, well-structured and convincing way).

  - This is how some of our PhD students received awards in recognition to the quality of their research. Have a look here: https://team.inria.fr/kerdata/awards/.

- **Additional complementary training** will be available, with the goal of preparing the PhD candidates for their postdoctoral career, should it be envisioned in academia, industry or in an entrepreneurial context, to create a startup company.

Mission confiée

Introduction

Large scale computing has demonstrated a profound impact on modern research and industries across many domains, enabling many breakthroughs through simulation and analytics. Today's “Big Data” projects generate datasets that are increasing exponentially in both complexity and volume, making their analysis, archival, and sharing one of the grand challenges of our modern society. In this context, an unprecedented opportunity is arising to automatically learn and gain insight from these massive amounts of complex data. Unsurprisingly, technologies such as deep learning have seen a rapid rise as a consequence. Major companies like Microsoft and Google changed their business strategies to become “AI-first” companies, aiming for supremacy in an area that can give them enormous competitive advantage if successful. Governments and national laboratories have acknowledged the profound implications and benefits of deep learning in accelerating science to solve the most challenging problems of society in various areas: climate, energy, healthcare, etc.

The rise of 5G and mobile technologies has shifted the predominantly centralized computational models used so far (clouds, data centers, supercomputers) towards more hybrid, loosely coupled models where centralized, high-performance computing infrastructures are combined with edge computing infrastructures of more modest capacity. Using this model, strategic computations can be placed on the entire path of the data from source (mobile sensors and devices) to the destination where it needs to be aggregated. By acting on streams of data directly, such near-data computations have two advantages: (1) they reduce the latency of obtaining the desired results; (2) they reduce resource utilization (I/O bandwidth involved in moving data, storage space on parallel file systems, core-hours needed on data centers, etc.). Unfortunately, this is not non-trivial for two reasons: (1) deep learning applications need to be distributed, which has non-trivial implications on the very nature of the learning process itself; (2) data sources (sensors, scientific instruments, mobile devices) are becoming mobile, which means they will constantly move away from their original entry point, therefore introducing the need to constantly migrate stateful computations close to them.

Thesis proposal

To address the aforementioned challenges, this thesis aims to build a deep learning framework specifically optimized to learn from mobile streams. Specifically, it will answer questions in two directions: (1) how to adapt current deep learning approaches such that they can use strategic computations on the path of data; (2) when and how to move stateful computations close to mobile data sources such that they can deliver low latency and high learning throughput. To this end, the thesis will start with a thorough analysis of the current challenges and opportunities in both directions. Specifically it will answer questions such as: how to learn directly from streams given the problem of critical forgetting (i.e., reinforcement of new patterns at the expense of older ones)? Can we design better streaming technologies that offer support to revisit and replay important old data? When is the data far enough so that it is worth migrating the computations closer to it? How can we capture the state of a deep learning training while it is in progress and move it to another destination efficiently? How to resume the training in an environment that looks different (e.g. different performance characteristics or resource constraints?) How can we link components such that migrated computations still play well with other computations that were left behind? Based on such questions and their answers, the goal is to design and develop algorithms and experimental prototype implementations that demonstrate the merit of such an approach in practice.

Principales activités

Enabling technologies

In the process of designing such a deep learning framework, we will leverage in particular techniques for data processing already investigated by the participating teams as proof-of-concept software, validated in real-life environments:

- The **VeloC** [1] project (Very Low Overhead Checkpointing System) is a multi-level checkpointing runtime for HPC supercomputing infrastructures and large-scale data centers sponsored by ECP (Exascale Computing Project). It aims to deliver high performance and scalability for complex heterogeneous storage hierarchies without sacrificing ease of use and flexibility. Checkpointing is well known as a fault-tolerance mechanism for tightly coupled HPC applications. Furthermore, it is an essential building block for many other use cases: suspend-resume, migration, debugging, revisit and reuse previous computational and data states, sharing of intermediate data for workflows that combine computations with analytics, maintain data provenance and history of changes to establish trust, reproduce results and explain datasets/establish correlations (particularly important for machine learning models)
- The **KerA** [2] approach for Cloud-based low-latency storage for stream processing (currently under development at Inria). By eliminating storage redundancies between data ingestion and storage, preliminary experiments with KerA successfully demonstrated its capability to increase throughput for stream processing.
- **Tensorflow/Keras** [3] is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.
- **Horovod** [4] is an open source framework that runs on top of the Tensorflow ecosystem and facilitates high-performance and scalability for data-parallel training by leveraging HPC communication patterns, notably all-reduce over MPI.

International visibility and mobility

The thesis will be co-supervised by **Bogdan Nicolae** from **Argonne National Laboratory** (ANL, USA), leveraging an existing collaboration framework between INRIA and ANL: the UNIFY Associate team and the JLESC international Laboratory.

Applicability

This PhD proposal has strong links with large scale deep learning applications, notably CANDLE [3] (Cancer Deep Learning Environment). CANDLE is a major effort based at Argonne that involves multiple partners and is part of ECP (Exascale Computing Project, the largest supercomputing initiative in the USA). It uses deep learning to solve three interconnected problems: (1) RAS pathway problem: guide multi-scale molecular dynamics (MD) runs through a large-scale state-space search, using unsupervised learning to determine the scope and scale of the next series of simulations based on the history of previous simulations; (2) drug response: use supervised machine learning methods to capture the complex, non-linear relationships between the properties of drugs and the properties of the tumors to predict response to treatment; (3) treatment strategy: use semi-supervised machine learning to automatically read and encode millions of clinical reports which will be used by the national cancer surveillance program to understand the broad impact of cancer treatment practices. CANDLE is one example of a real-life application that can benefit from the proposal: data could be collected from hospitals and learned from in real time, instead of being centralized and analyzed in an offline fashion on supercomputing architectures (as it is today).

References

[1] VeloC: Towards High Performance Adaptive Asynchronous Checkpointing at Large Scale. Bogdan Nicolae, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, Franck Cappello. IPDPS 2019: 911-920

[2] O.C. Marcu, A. Costan, G. Antoniu, M. Pérez-Hernández, B. Nicolae, et al.. “KerA: Scalable Data Ingestion for Stream Processing”. ICDCS 2018 – 38th IEEE International Conference on Distributed Computing Systems, Vienna, Austria, pp.1480-1485, 2018

[3] https://www.tensorflow.org/

[4] https://eng.uber.com/horovod/

Compétences

- An excellent Master degree in computer science or equivalent
- Strong knowledge of computer networks and distributed systems
- Basic understanding of (distributed) file systems and storage solutions

- Ability and motivation to conduct high-quality research, including publishing the results in relevant venues
- Strong programming skills (e.g. C/C++, Python).
- Working experience in the areas of Big Data management, Cloud computing, HPC, is an advantage
- Very good communication skills in oral and written English.
- Open-mindedness, strong integration skills and team spirit

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement

Rémunération

Rémunération mensuelle brute de 1982 euros les deux premières années et 2085 euros la troisième année

Informations générales

- **Thème/Domaine :** Calcul distribué et à haute performance
  Système & réseaux (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2020-10-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2020-06-30

Contacts

- **Équipe Inria :** KERDATA
- **Directeur de thèse :**
  Antoniu Gabriel / 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 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 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 200 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
- Strong knowledge of computer networks and distributed systems
- Basic understanding of (distributed) file systems and storage solutions
- Ability and motivation to conduct high-quality research, including publishing the results in relevant venues
- Strong programming skills (e.g. C/C++, Python).
- Working experience in the areas of Big Data management, Cloud computing, HPC, is an advantage
- Very good communication skills in oral and written English.
- Open-mindedness, strong integration skills and team spirit

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Merci de déposer en ligne CV, lettre de motivation et éventuelles recommandations

Pour plus d'information, contactez gabriel.antoniu@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 :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.