Offre n°2025-08839

Doctorant F/H Optimisation de l'inférence de l'intelligence artificielle à grande échelle dans des environnements distribués dans le Cloud

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

Le centre Inria de l'université de Bordeaux est un des neuf centres d'Inria en France et compte une vingtaine 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

Ce sujet de thèse est proposé dans le cadre d'une collaboration entre Inria et hive (https://www.hivenet.com). hive fournit Hivenet, une infrastructure Cloud entièrement distribuée basée sur les machines des membres de la communauté. En d'autres termes, les membres de Hivenet partagent leurs propres ressources informatiques et de stockage dans le Cloud, ce qui profite ensuite aux autres utilisateurs et utilisatrices. L'un des avantages de cette architecture est qu'elle a un

impact positif sur les émissions de carbone en utilisant les ressources existantes au lieu de construire de nouveaux centres de données.

Nous nous concentrons sur l'émergence de nouvelles applications qui s'adaptent naturellement à l'exécution distribuée, telles que l'inférence de modèles d'intelligence artificielle (IA) à grande échelle qui nécessitent l'exécution d'un grand nombre de tâches indépendantes. Ces applications sont bien adaptées à l'exécution distribuée sur des ressources informatiques interconnectées, même si le réseau est nettement moins puissant que celui d'un superordinateur.

Dans le contexte de l'inférence de l'IA à grande échelle, plusieurs problèmes doivent être résolus :

- Chaque génération de token doit a priori passer par plusieurs unités de traitement graphique (GPUs), chacune stockant différentes parties du modèle. Il y a un problème de partitionnement du modèle [1][2] et aussi un problème associé à la construction de chemins d'inférence pour minimiser la latence (faire des groupes de ressources proches) ;
- La quantité d'inférence requise variera naturellement dans le temps, et l'ensemble des ressources mises à la disposition du calcul variera également. Il existe donc un problème de planification statique (décider quelles ressources sont susceptibles de participer et y stocker les modèles) et un problème dynamique (comment allouer les nouvelles demandes) ;
- En fonction des modèles utilisés, certaines tâches d'inférence sont naturellement placées sur certaines ressources (par exemple, parce que les tokens précédents ont été générés dans cette ressource). En termes de tolérance aux défaillances, le calcul peut facilement être redémarré (en sachant quels tokens ont été générés), mais à un coût élevé. Cela pose des problèmes d'allocation des ressources en fonction des statistiques de disponibilité des ressources.

Mission confiée

L'objectif de cette thèse est de proposer et d'adapter des techniques pour améliorer la latence, le débit ou l'utilisation des ressources de l'inférence IA à grande échelle dans un environnement Cloud entièrement distribué.

Le travail se concentrera sur les modèles complexes (tels que les LLMs) qui nécessitent plusieurs GPU en raison de leurs besoins en mémoire. Ces GPUs constitueront également une plateforme informatique complexe, compte tenu de leur hétérogénéité potentielle, de leur volatilité et de leur répartition géographique dans l'infrastructure du Cloud.

Les techniques d'optimisation proposées se concentreront sur les méthodes d'ordonnancement et d'allocation des ressources afin d'optimiser le temps d'exécution des tâches d'inférence et leur communication.

Un objectif de cette collaboration avec hive est de déployer les modèles optimisés à grande échelle dans des nœuds de calcul :

- du même fournisseur (avec des GPU homogènes de niveau desktop disponibles à proximité les uns des autres) ; et
- de différents membres de la communauté (où la latence de communication entre les fournisseurs doit être prise en compte, et où les modèles de GPU sont plus hétérogènes).

Dans le contexte de ce sujet, nous considérons également qu'il est pertinent de mentionner ses similitudes avec le traitement des flux de données (DSP). Plus précisément, les DSPs sont des modèles représentant le traitement continu des données (apparenté à la génération de tokens), qui passent par divers opérateurs (répliqués si nécessaire) représentés par un DAG. Il existe une littérature abondante sur le placement des opérateurs dans de tels modèles, avec l'optimisation de diverses mesures [3] (latence, débit, communications, mise à l'échelle, tolérance aux pannes, etc.), bien que la plupart d'entre elles concernent le cas de l'exécution dans des nuages [4] (moins de contraintes de ressources et moins d'hétérogénéité).

Principales activités

- Recherche bibliographique (littérature sur l'ordonnancement et l'IA)
- Modélisation de l'inférence en IA (pour son optimisation)
- Conception, programmation et validation d'algorithmes
- Intégration de composants ou de modules logiciels dans des cadres logiciels existants
- Publication des résultats de la recherche dans des articles scientifiques
- Présentation et diffusion des résultats lors de conférences et de workshops

Activités complémentaires :

- Participation aux cours obligatoires de l'école doctorale (EDMI - Université de Bordeaux)
- Participation facultative à l'encadrement des étudiants en stage de master
- Enseignement facultatif à l'université de Bordeaux ou à l'école d'ingénieurs ENSEIRB-MATMECA

Compétences

- Connaissance intermédiaire du calcul à haute performance et de l'informatique en nuage ;
- Connaissance débutante des algorithmes et des problèmes d'optimisation (une connaissance plus approfondie est un atout) ;
- Une bonne connaissance des LLMs sera appréciée ;
- Bon niveau de développement de logiciels sous des systèmes d'exploitation de type UNIX ;

- Bonnes aptitudes rédactionnelles ;
- Volonté de travailler dans un environnement diversifié et international.

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Rémunération

La rémunération sera de 2200€ brut, puis de 2300€ brut à partir de 2026 (avant charges et taxes)

Informations générales

- **Thème/Domaine :** Calcul distribué et à haute performance
  Système & réseaux (BAP E)
- **Ville :** Talence
- **Centre Inria :** Centre Inria de l'université de Bordeaux
- **Date de prise de fonction souhaitée :** 2025-10-01
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2025-06-30

Contacts

- **Équipe Inria :** TOPAL
- **Directeur de thèse :**
  Lima Pilla Laercio / laercio.lima@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'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

- Goût pour la conception d'algorithmes et les problèmes d'optimisation ;
- Curiosité et appétence pour l'exploration et l'expérimentation ;
- Ouverture au travail en équipe ;
- Motivation pour la mise en œuvre de pratiques scientifiques éthiques et rigoureuses.

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

Si vous êtes intéressés par cette offre, merci de bien vouloir candidater via le site jobs.inria avec les documents suivants :

- cv
- lettre de motivation

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.