Offer #2025-08886

Doctorant F/H Intégration de l’ordonnancement asynchrone des communications réseau et de l’ordonnancement des tâches

The offer description below is in French

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

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…

Context

Cette thèse s'inscrit dans le cadre du projet Exa-Soft du programme Numpex.

Dans le cadre du calcul haute-performance, les machines sont désormais très hétérogènes, équipées d’accélérateurs tels que les GPU ou les FPGA. Ces différentes unités se programment avec des paradigmes différents, et

pour ajouter à la complexité, il faut gérer les transferts de données entre les unités, l'ordonnancement, et l'équilibrage de charge.

Pour tirer profit d'une telle plate-forme, l'équipe STORM a proposé le support d'exécution StarPU qui gère ces problématiques de manière générique et indépendante de l'application. Pour des grappes de calcul – plusieurs noeuds interconnectés par un réseau d'interconnexion – StarPU confie l'exploitation du réseau à une bibliothèque de communication qui implémente l'interface MPI, d'utilisation standard en HPC.

Par ailleurs, l'équipe TADAAM développe la bibliothèque de communication NewMadeleine, dont l'originalité est d'appliquer une stratégie d'optimisation à la volée sur les flux de communications issus des différents threads, en assurant une progression asynchrone en tâche de fond. Elle repose essentiellement sur le principe de programmation événementielle et de messages actifs, ce qui permet le déroulement des communications sans intervention de l'application.

Les besoins de StarPU en terme de communications ne sont pas les mêmes que ceux d'une application MPI classique, notamment en matière d'irrégularité, de réactivité, de multi-threading, de nombre de requêtes actives simultanément, alors qu'ils correspondent parfaitement au cahier des charges de NewMadeleine. Un portage se StarPU sur l'interface native de NewMadeleine a été réalisé pour en exploiter au mieux les propriétés de progression, de passage à l'échelle [6], ainsi qu'une intégration spécifique [7] des opérations collectives.

Assignment

Objectif

L'objectif de cette thèse est de renforcer le dialogue entre NewMadeleine et StarPU dans la gestion des communications. Il s'agit d'exploiter la connaissance du futur que l'on peut extraire du graphe de tâches de façon à optimiser les communications.

Responsables:

• Alexandre DENIS (Inria, Alexandre.Denis@inria.fr),
• Samuel THIBAULT (U. Bordeaux, Samuel.Thibault@inria.fr)
• Philippe SWARTVAGHER (Bordeaux INP, Philiipe.Swartvagher@inria.fr)

Commentaires

Les développements à réaliser se feront dans les bibliothèques StarPU et NewMadeleine disponibles en open source et se font en langage C. Il serait

souhaitable qu'ils soient réalisés par une personne à l'aise en programmation réseau et système.

References

[1] Emmanuel Agullo, Olivier Aumage, Mathieu Faverge, Nathalie Furmento, Florent Pruvost, Marc Sergent, and Samuel Thibault. Achieving High Performance on Supercomputers with a Sequential Task-based Programming Model. IEEE Transactions on Parallel and Distributed Systems, 2017.
[2] Cédric Augonnet, Samuel Thibault, Raymond Namyst, and Pierre-André Wacrenier. StarPU: A Unified Platform for Task Scheduling on Heterogeneous Multicore Architectures. Concurrency and Computation: Practice and Experience, Special Issue: Euro-Par 2009, 23:187–198, February 2011.
[3] Olivier Aumage, Elisabeth Brunet, Nathalie Furmento, and Raymond Namyst. NewMadeleine: a Fast Communication Scheduling Engine for High Performance Networks. In Workshop on Communication Architecture for Clusters (CAC 2007), workshop held in conjunction with IPDPS 2007, Long Beach, California, United States, March 2007.
[4] Guillaume Beauchamp. Portage de StarPU sur la bibliothèque de communication NewMadeleine. Master's thesis, Universite Bordeaux, September 2017.
[5] Alexandre Denis. pioman: a pthread-based Multithreaded Communication Engine. In Euromicro International Conference on Parallel, Distributed and Network-based Processing, Turku, Finland, March 2015.
[6] Alexandre Denis. Scalability of the NewMadeleine Communication Library for Large Numbers of MPI Point-to-Point Requests. In CCGrid 2019 - 19th Annual IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing, Larnaca, Cyprus, May 2019.
[7] Alexandre Denis, Emmanuel Jeannot, Philippe Swartvagher, and Samuel Thibault. Using Dynamic Broadcasts to improve Task-Based
4
Runtime Performances. In Euro-Par 2020, Euro-Par 2020, Warsaw, Poland, August 2020. Rzadca and Malawski, Springer.
[8] Alexandre Denis and François Trahay. MPI Overlap: Benchmark and Analysis. In International Conference on Parallel Processing, 45th International Conference on Parallel Processing, Philadelphia, United States, August 2016.

Main activities

Détails de l'activité :

Enregistrement de la mémoire. Les réseaux haute performance tels qu'InfiniBand sont programmés directement depuis l'espace utilisateur.

Pour que la carte puisse accéder directement aux données de l'utilisateur, en mémoire virtuelle, il faut au préalable enregistrer les pages mémoire pour qu'elle connaisse leur adresse physique. Cette opération est appelée enregistrement mémoire. Cette opération est coûteuse. Dans les applications HPC classiques, son coût est habituellement amorti au travers d'un cache d'enregistrement, car les buffers sont souvent re-utilisés. En revanche, dans le cadre de StarPU, la réception a souvent lieu dans un buffer alloué dynamiquement, utilisé une seule fois, déjouant les politiques reposant sur un cache.

Grâce a la connaissance du futur dans le graphe de tâche, il est possible d'effectuer un enregistrement à la volée des zones mémoire et d'amortir son coût en l'anticipant. En effet, le moment où une zone mémoire sera utilisée pour une émission ou une réception réseau est prévisible. Ces travaux font l'objet actuellement d'un stage.

Il restera à étudier le coût résiduel de l'enregistrement mémoire, étudier ses interaction avec le GPU, et notamment l'enregistrement mémoire demandé par le GPU, les interactions entre l'allocation dynamique de buffers pour une réception réseau et le placement souhaité par l'ordonnanceur de tâche.

Priorités des communications. StarPU attache des priorités aux tâches, NewMadeleine est capable d'ordonnancer les paquets en fonction de leurs priorités. La stratégie actuelle attache aux paquets la priorité associée à leur tâche dans le graphe de tâche. Il n'est toutefois pas certain que ce soit la solution optimale.

Par ailleurs, l'ordonnancement des paquets par priorités dans un contexte HPC n'est pas encore totalement compris: tenir compte des priorités relatives entre destinataires différents, étudier les interactions entre priorités et protocole de rendez-vous, tenir compte des priorités dans la construction des arbres de diffusion, etc.

Il est également possible d'imaginer prendre en compte les communications futures, que l'on peut déduire du graphe de tâche, dans l'ordonnancement des communications. Par exemple, si un paquet très prioritaire devra prochainement être envoyé, il est envisageable de prendre la décision de ne pas monopoliser le réseau avec un envoi moins prioritaire,même s'il est déjà prêt à envoyer.

Adaptation dynamique à la pression des communications. Il est possible pour certaines applications d'adapter dynamiquement la granularité des tâches, ce qui a un impact sur le schéma de communication. Nous pouvons imaginer que NewMadeleine puisse tenir StarPU informé de la pression sur le

système de communications (en fonction de la taille du backlog de communications en attente, par exemple), de façon à adapter dynamiquement la granularité des calculs.

Types de données. Le type des données, au sens du datatype MPI, influe sur la façon de les gérer et les performance obtenues. La différence se fait essentiellement selon que les données sont contigües en mémoire ou non, ainsi que la taille des fragments.

Les différentes méthodes de communications ne présentent pas les mêmes caractéristiques et donc pas les mêmes performances selon le type de données. On pourra envisager de choisir dynamiquement la méthode de communication en fonction du type des données côté émetteur et côté récepteur.

Skills

Compétences techniques et niveau requis : langage C, développement système

Langues : anglais courant

Benefits package

- 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

Remuneration

Le salaire sera de 2200€ brut puis 2300€ à partir de 2026 (avant prélèvements sociaux et taxes).

General Information

- **Theme/Domain :** Distributed and High Performance Computing
  Scientific computing (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 :** TADAAM
- **PhD Supervisor :**
  Denis Alexandre / Alexandre.Denis@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

- curiosité
- autonomie
- rigueur

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

Si vous êtes intéressés, merci de candidater via le site jobs.inria avec les documents suivants :

- cv
- lettre de motivation

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.