Offer #2022-04380

Ingénieur Scientifique - Passage à l'échelle des solveurs d'EDO pour la biologie computationnelle

The offer description below is in French

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Other valued qualifications : Thése/PhD

Fonction : Temporary scientific engineer

Corps d'accueil : Ingénieur de Recherche (IR)

Level of experience : From 3 to 5 years

About the research centre or Inria department

Le centre de recherche Inria de Lyon (précédemment antenne lyonnaise du centre Inria de Grenoble), est le 9ème centre de recherche Inria, créé formellement en décembre 2021.

Il regroupe environ 270 personnes (dont 110 salariés Inria) au sein de 15 équipes de recherche et de services supports à la recherche.

Ses effectifs sont distribués à ce stade sur 2 campus : à Villeurbanne, La Doua (Centre / INSA Lyon / UCBL) d'une part, et à Lyon Gerland (ENS de Lyon) d'autre part.

Une 3ème implantation devrait voir le jour dans le courant de 2022. Les équipes sont essentiellement hébergées chez nos partenaires.

Les équipes du centre travaillent en lien étroit avec les établissements de recherche et d'enseignement supérieur (ENS de Lyon, UCBL, INSA Lyon, ...), leurs laboratoires, et autres organismes de recherche de Lyon (CNRS, INRAE, pôles de compétitivité, ...), mais aussi avec les acteurs économiques lyonnais et régionaux.

De nombreuses collaborations sont par ailleurs en cours à l'international. Le centre de Lyon est présent dans les domaines du logiciel, du calcul distribué et haute performance, des systèmes embarqués, du calcul quantique et de respect de la vie privée dans le monde numérique, mais aussi de la santé et de la biologie numériques.

Context

Résumé

En biologie, la grande majorité des systèmes peut être modélisée sous la forme d'équations différentielles ordinaires (ODE). Modéliser plus finement des objets biologiques mène à augmenter le nombre d'équations. Simuler des systèmes toujours plus grands mène également à augmenter le nombre d'équations. Par conséquent, nous observons une explosion de la taille des systèmes d'ODE à résoudre. Un verrou majeur est la limitation des logiciels de résolutions numériques d'ODE (solveur ODE) à quelques milliers d'équations à cause de temps de calcul prohibitif. L'AEx ExODE s'attaque à ce verrou via 1) l'introduction de nouvelles méthodes numériques qui tireront parti de la précision mixte qui mélange plusieurs précisions de nombre flottant au sein d'un schéma de calcul, 2) l'adaptation de ces nouvelles méthodes pour des machines de calcul de prochaines générations qui sont fortement hiérarchiques et hétérogénes et composées d'un grand nombre de CPUs et GPUs. Depuis un an, une nouvelle approche du Deep Learning se propose de remplacer les Recurrent Neural Network (RNN) par des systèmes d'ODE. Les méthodes numériques et parallèles d'ExODE seront évalué et adapté dans ce cadre afin de permettre l'amélioration de la performance et de l'exactitude de ces nouvelles approches.

Description du projet

Les équations différentielles ordinaires (EDO) sont un formalisme de modélisation majeur en biologie (voir ci-dessous pour plus de détails). Ces exemples se traduisent souvent en des systèmes de N équations fortement connectées, ce qui donne des temps ce calcul en $O(N^2)$:
y0(t) = f (t, y(t)), t > 0, y(0) = y0, y(t) RN , f : R × RN RN .

Ces modèles sont souvent limités à quelques milliers d'équations à cause des temps de calcul et de l'utilisation mémoire nécessaire pour leur résolution. La capacité des logiciels de résolution d'EDO (solveur EDO) à résoudre dans un temps raisonnable devient critique de part le besoin de passage à l'échelle i.e. une augmentation importante de la taille des systèmes modélisés : 100 à 1000 fois plus d'équations (soit une augmentation de la complexité de 10^4 à 10^6).

Les solveurs EDO ont des caractéristiques de précisions et de stabilité. Un schéma peut être stable, mais pas précis (par exemple Euler implicite) ou précis mais pas stable (Runge-Kutta d'ordre élevés). Pour les applications biologiques envisagées, le critère de stabilité est plus important que la précision, ce qui tendrait à utiliser des schémas d'ordres peu élevés.

Afin de rendre toujours plus performante l'utilisation de l'apprentissage profond, la précision mixte ainsi que des schémas de calcul les utilisant (e.g. convolution) ont été introduites dans les architectures matérielles depuis plusieurs années e.g., Intel VNNI, Intel Nervana, NVidia Tensor-Core, Google TPU. L'idée de la précision mixte est de combiner l'utilisation de demi-précision (16 bits) et de simple précision (32 bits) afin de réduire l'utilisation mémoire et augmenter la densité de calcul. Si le matériel a été développé pour l'apprentissage profond classique, l'utilisation de précision mixte est aussi considérée en algèbre linéaire [12, 4] ou pour la résolution de problèmes linéaires qui surviennent dans la discrétisation d'équations au dérivée partielles [2]. L'impact de la précision sur le temps de calcul et la stabilité des résultats a été étudié dans le cadre des ODEs [13]. Mais ces études pour des méthodes mélangeant plusieurs précisions n'ont été faite que pour des méthode d'ordre élevé et des tailles de systèmes relativement petites [11].

Bien entendu, le passage en précision mixte peut limiter la précision des solveurs EDO actuels, et possiblement induire des instabilités numériques. Cependant, les chercheurs utilisent en pratique les paramètres par défaut des logiciels de résolution d'EDO, avec des précisions de l'ordre de 3 à 6 décimales (par exemple les packages deSolve de R ou scipy.integrate.ode en python se basent sur ODEPACK [6]). Avec des flottants en semi-précision, qui ont 3-4 décimales de précision, cette précision est largement atteignable. Les méthodes prédicteur-correcteur sont bien adaptées aux calculs en précision mixte, avec le calcul d'une prédiction réduite et de la correction en précision étendue. Il sera aussi utile tester ces méthodes en précision réduite lorsque l'on ne s'intéressera qu'aux propriétés statistiques des solutions. De plus, un des aspects exploratoires de ce projet est de mélanger plusieurs précisions au sein même d'un calcul (e.g. entrées en demi-précision et accumulateur en précision simple) afin de limiter le recours à des méthodes trop coûteuse de prédiction-correction.

Afin d'exploiter au maximum les ressources des plate-formes de calcul, il est nécessaire de prendre en compte la parallélisation des solveurs d'ODE. C'est un domaine étudié depuis de nombreuses années [15, 7], il existe 3 grands axes de parallélisation : méthode, temps et système. La plupart des approches se sont focalisés sur les 2 premiers car ils s'appliquent à des systèmes de relativement petite taille. Les systèmes que nous considérons sont de grande taille et donc permettent d'exposer suffisamment de parallélisme au niveau système. Par conséquence, les méthodes d'optimisation d'EDO tel que la transformation de boucle [14, 10] ne sont donc pas pertinentes dans notre cas. Il est à noter que les différentes précisions, méthodes de résolution, de corrections et de parallélisme amènent des compromis performance et précision qui ne seront pas les mêmes suivant les systèmes. Il est donc important d'étudier ces compromis et de proposer des solutions automatisant au maximum ce choix. Plusieurs méthodes [9, 8] ont été proposées dans ce cadre mais en ne prenant pas en compte la demi-précision/correction. De plus les solutions proposées sont généralistes alors que nous proposons d'étudier un domaine d'application précis en intégrant leurs spécificités dès la conception

Pour résumer, l'objectif de ce poste est d'intégrer de nouvelles approches de précisions mixtes dans la résolution d'EDO sur des plate-formes de calcul modernes (et futures) pour la biologie computationnelle (et potentiellement pour l'apprentissage profond à base d'EDO, voir ci-dessous pour plus d'explications). Ajoutons qu'au-delà des cas d'utilisation en biologie qui sont au coeur de ce projet, la résolution d'EDO est un verrou technologique majeur pour d'autres modèles en biologie. Au-delà de la biologie, la résolution d'EDO est au coeur d'un nouveau courant de l'apprentissage profond afin de pouvoir répondre aux limitations des méthodes actuelles [1]. En effet, les mécanismes utilisés actuellement discrétisent les espaces de recherche, or dans le cas où les données utilisées se basent sur des variables continues, l'utilisabilité de ces méthodes est sévèrement limitée. Dans [1], les auteurs proposent de remplacer la méthode discrète qui est au coeur du deep learning actuel i.e. les Recurrent Neural Network (RNN) qui sont les briques de base des convolution neural networks (CNN) [5] qui sont la pierre angulaire de l'explosion actuelle de l'utilisabilité du deep learning. Les auteurs de [1] ainsi que plusieurs travaux récents [3] proposent de remplacer les RNNs par des méthodes à base d'EDO qui permettent de prendre en compte ces variables continues, d'apprendre de meilleur modèles mais aussi et surtout qui permettent de se baser sur plusieurs dizaines d'années de recherche théorique sur les EDO. Or, pour être efficace, ces approches nécessitent la résolution d'un nombre très important de systèmes d'EDO de grande taille (de plusieurs centaines de milliers à plusieurs dizaines de millions de paramètres). Il est donc nécessaire de disposer de mécanisme performant de résolution d'EDO. En se basant sur ces similarités, nous explorerons la possibilité que les approches développées pour la biologie computationelle pourront avoir un impact majeur sur ces nouvelles approches de deep learning aussi bien en terme de performance et précision que passage à l'échelle de ces méthodes.

Références

[1] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In Advances in Neural Information Processing Systems, pages 6571-6583, 2018.

[2] Dominik Göddeke, Robert Strzodka, and Stefan Turek. Performance and accuracy of hardware-

oriented native-, emulated-and mixed-precision solvers in fem simulations. International Journal of Parallel, Emergent and
Distributed Systems, 22(4) :221-256, 2007.

[3] Will Grathwohl, Ricky TQ Chen, Jesse Betterncourt, Ilya Sutskever, and David Duvenaud. Ffjord : Free-form
continuous dynamics for scalable reversible generative models. arXiv preprint arXiv :1810.01367, 2018.

[4] Azzam Haidar, Stanimire Tomov, Jack Dongarra, and Nicholas J. Higham. Harnessing gpu tensor cores for fast
fp16 arithmetic to speed up mixed-precision iterative refinement solvers. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC '18, pages 47 :1-47 :11,
Piscataway, NJ, USA, 2018. IEEE Press.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In The
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[6] Alan C Hindmarsh. Serial fortran solvers for ode initial value problems. URL : https://computation.llnl.gov/casc/odepack/ [cited June 6, 2019], 2002.

[7] P.J. Van Der Houwen and B.P. Sommeijer. Parallel iteration of high-order runge-kutta methods with stepsize
control. Journal of Computational and Applied Mathematics, 29(1) :111-127, 1990.

[8] Natalia Kalinnik, Matthias Korch, and Thomas Rauber. An efficient time-step-based self-adaptive algorithm
for predictor-corrector methods of runge-kutta type. Journal of Computational and Applied Mathematics, 236(3) : 394 - 410, 2011. Aspects of Numerical Algorithms, Parallelization and Applications.

[9] Natalia Kalinnik, Matthias Korch, and Thomas Rauber. Online auto-tuning for the time-step-based parallel
solution of odes on shared-memory systems. Journal of Parallel and Distributed Computing, 74(8) :2722-2744, 2014.

[10] Matthias Korch and Tim Werner. Accelerating explicit ode methods on gpus by kernel fusion. Concurrency and
Computation : Practice and Experience, 30(18) :e4470, 2018. e4470 cpe.4470.

[11] Tomonori Kouya. Practical implementation of high-order multiple precision fully implicit runge-kutta methods
with step size control using embedded formula. arXiv preprint arXiv :1306.2392, 2013.

[12] Xiaoye S. Li, James W. Demmel, David H. Bailey, Greg Henry, Yozo Hida, Jimmy Iskandar, William Kahan,
Suh Y. Kang, Anil Kapur, Michael C. Martin, Brandon J. Thompson, Teresa Tung, and Daniel J. Yoo. Design, implementation and testing of extended and mixed precision blas. ACM Trans. Math. Softw., 28(2) :152-205, June 2002.

[13] L. Murray. Gpu acceleration of runge-kutta integrators. IEEE Transactions on Parallel and Distributed Systems,
23(1) :94-101, Jan 2012.

[14] T. Rauber and G. Rünger. How do loop transformations affect the energy consumption of multi-threaded runge-kutta methods ? In 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based
Processing (PDP), pages 499-507, March 2018.

[15] Thomas Rauber and Gudula Rünger. Parallel execution of embedded and iterated runge-kutta methods. Concurrency : Practice and Experience, 11(7) :367-385, 1999.

Assignment

L'ingénieur aura pour mission de contribuer aux activités de développement et d'intégration des recherches de l'action exploratoire ExODE (Scaling the solving of Ordinary Differential Equation for Computational Biology) sur les sujets suivants: méthode numérique et parallèle pour la résolution d'équation différentielle ordinaire (ODE), schéma de résolution d'ODE en précision mixte, adaptation de méthode numérique pour la résolution d'ODE appliquée à la biologie computationnelle (et le deep learning) et plus généralement de participer à l'ensemble des activités d'ingénieurie de l'action exploratoire ExODE.

Main activities

En collaboration avec les chercheurs impliqués dans l'AEx ExODE, l'ingénieur sera responsable du développement des différentes méthodes de résolutions numériques proposés. Des premiers prototypes ont été développé, l'ingénieur aura pour mission de les rendre plus robuste, utilisable et évoluable (faciliter l'ajout/extension de méthodes et de modèles biologiques). L'ingénieur sera impliqué dans la formation de modélisateurs et d'autres ingénieurs sur l'utilisation et le développement au sein de la solution logicielle qu'il aura développé. Suivant l'appétence du candidat, une partie d'implication dans la recherche et l'écriture d'article est possible.

Skills

Une bonne compétence en C++ est essentielle. Des connaissances en CUDA, Kokkos, vectorization, arithmétique en précision mixte seront des plus. Une maitrise de l'anglais est nécessaire et celle du français un plus mais non essentielle.

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

Selon grille indiciaire et selon profil.

General Information

- **Theme/Domain :** Computational Biology
  Scientific computing (BAP E)
- **Town/city :** Villeurbanne
- **Inria Center :** Centre Inria de Lyon
- **Starting date :** 2022-09-01
- **Duration of contract :** 1 year, 7 months
- **Deadline to apply :** 2022-07-31

Contacts

- **Inria Team :** BEAGLE
- **Recruiter :**
  Rouzaud-cornabas Jonathan / jonathan.rouzaud-cornabas@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

L'ingénieur recruté sera en intéraction forte avec des mathématiciens, des modélisateurs (en biologie et santé) et des informaticiens (HPC, arithmétique). Une forte appétence pour le travail inter/trans-disciplinaire est essentielle. Une expertise dans un des domaines est demandé:

- schéma de résolution numérique pour les équations différentielles ordinaires
- algorithmes de parallélisation à grande échelle (cluster avec machines hybrides CPU-GPU) et les solutions actuelles de parallélisation (OpenMP, Kokkos, CUDA)
- arithmétique en précision mixte

Une envie d'apprendre des compétences des autres domaines est demandée.

Instruction to apply

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.