PROPOSITION SUJET DE THESE TTI.5

Titre de la thèse : **Carbon-aware High Performance Artificial Intelligence**

Equipe de direction

Directeur:
NOM Prénom : **Claude TADONKI, PhD, HDR**
Spécialité / domaine : **Calcul Haute Performance, compilation, Compilation, Recherche Opérationelle**
Centre et département de rattachement : **Centre de Recherche en Informatique (CRI), Math et Systèmes**

Co-directeur:
NOM Prénom : **Petr DOKLADAL, PhD, HDR**
Spécialité / domaine : **Intelligence Artificielle, Analyse d'images, Morphologie Mathématique**
Centre et département de rattachement : **Centre de Morphologie Mathématique (CMM), Math et Systèmes**

Co-directeur:
NOM Prénom : **Youssef MESRI, PhD, HDR**
Spécialité / domaine : **Mathématiques Appliquées, Calcul Haute Performance, Mécanique des Fluides Numérique**
Centre et département de rattachement : **Centre de Mise en Forme de Matériaux (CEMEF), Mécanique et Matériaux**

Département et Ecole Doctorale de rattachement : Mathématiques et Systèmes, École doctorale 621, Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Description du sujet (2 pages maximum, références bibliographiques incluses) :

Artificial Intelligence (AI) is becoming ubiquitous in modern life, with various applications of genuine interest in uncountable domains of everyday life. It enables computers to perform "intelligent" tasks such as decision making, problem solving, perception/identification and even understanding human communication and behavior and predict his needs and actions.

Current and future expectations from AI are challenging, yet significant advances have already been made thanks to powerful computational facilities and exascale datasets that are available to foster and unleash AI capabilities. This comes, however, in pair with the increase of energetic consumption and carbon impact.

Considering the importance of AI, it follows that there is a clear benefit of optimizing AI implementations. The aim of this PhD is to foster an energy-aware design of AI technology. The focus will be given to two mainstream computing platforms: standard HPC solutions (clusters, and datacenters) and embedded devices. In both, the consumed energy is critical. In the former, the available energy is considered unlimited, thus the main concern in this aspect is generally on the associated cost (electricity and cooling system) and the carbon footprint [3]. In the latter, energy consumption is taken into account at the design time since the targeted devices are likely to be battery powered. Nonetheless, the battery life-cycle's ecological burden is non-negligible too.

We point out the following objectives:

A) In AI, the training step is computationally intensive (the training process is iterative and slows down when the dataset is large). Thus, implementing a training process on a computing device with a limited energy budget is a challenging task that has never been considered. Other kinds of limitations like memory space or specific data types are to be considered too with large-scale training. Formal approaches can be found in [1, 2] and energy-aware design and an exhaustive evaluation at runtime can be found in [10].

An example of devices allowing to reduce the energy consumption is the Intel Neural Compute Stick (NC), an energy efficient chip designed to accelerate the execution of Deep Learning algorithms. The device comes with frameworks like «OpenVINO». Even though the NC Stick is energy efficient, its usage is limited to the execution of low-power algorithms - it has not (yet) been considered for training. Thus, we aim at optimizing AI models accordingly by reducing their space-time complexity as well as data movement. We seek a methodology that will address such specific devices and other relevant architectures too for an optimal deployment.

B) While the energetic budget of an AI model is complex to evaluate, one of the key factors is the energy budget of atomic operation (see [4] for a more general study of this approach). Hence, the energy of both training and application can be reduced by using a lower data precision. In some cases, this stands as a constraint because the device only operates with specific data types, in other cases it is a choice guided by optimization purposes. The most popular approach in this context is the so-called «mixed precision [9]», where the core of the algorithm is implemented with a lower precision and the global correction with a higher precision. This technique is common in numerical computation, especially with iterative schemes (like the *Conjugate Gradient*, *GMRES*, ⋯). The objective here is to deeply investigate an implementation methodology that acts on the data types for the purpose of energy optimization.

This PhD will tackle two notoriously costly AI applications, namely :

1) Real-time machine vision applications, using deep-learning techniques. Machine vision

applications have recently equaled and even exceeded the performance of the human eye in number of areas. This comes at the price of using large-scale datasets and large models. While we still cannot achieve data frugality without sacrificing the performance, one solution consists of reducing the energy budget of the computations.

2) Fluid flow prediction using Physics Informed Neural Networks [5, 6]. The *Physics Informed Neural Networks* approach is one of the emerging and promising Deep Learning approaches to predict physical phenomena modelled by physical equations such as those governing fluid flow dynamics. Indeed, the prior knowledge from the governing equations of the physical model can be embedded as a constraint in the deep learning loss function to better generalize and optimize the learning process in physically conservative spaces. The input datasets are generated from a parallel flow solver running on CPU based supercomputers [8] and the machine learning is performed on GPUs using Tensorflow. The cost of generating and training datasets is very high even using parallel computing systems. The common and main cost of both data generator (Fluid Solver) and Deep Learning Optimizer is mainly related to linear algebra operations i.e. *large scale linear systems to be solved for the Fluid Solver and the gradient descent based optimization for the DL Optimizer*. This PhD proposal will investigate new approaches to significantly reduce the complexity of such operations using matrix compression as in [7] and also mixed precision [9]. Note that reducing the complexity of our algorithms involves controlling the associated approximation errors. A compromise must be found between accuracy and low-cost solutions.

The research methodology for this PhD will follow three main axes : *computer science* (optimization of AI models and efficient implementations) ; *applied mathematics* (formal transformations, mixed precision, iterative solvers, numerical optimization); *applications* (computer vision and fluid mechanics).

References:

1. **Claude Tadonki**, Mitali Singh, Jose Rolim and Viktor K. Prasanna, *Combinatorial Techniques for Memory Power State Scheduling in Energy Constrained Systems*, Workshop on Approximation and Online Algorithms (WAOA), WAOA2003 (LNCS/Springer), Budapest, Hungary, September 2003 .
2. **Claude Tadonki** and Jose Rolim, *An analytical model for energy minim*ization, III Workshop on Efficient and Experimental Algorithms, WEA04 (LNCS/Springer), Angra dos Reis, Brazil, May 2004.
3. Ferreira Leite, A. Boukerche, A. C. Magalhaes Alves de Melo, C. Eisenbeis, **C. Tadonki**, and C. Ghedini Ralha, *Power-Aware Server Consolidation for Federated Clouds*, J Concurrency and Computation: Practice and Experience (CCPE), ISSN: 1532-0626, Wiley Press, New York, USA., 2016.
4. Alessandro Ferreira Leite, **Claude Tadonki**, Christine Eisenbeis, Alba de Melo, *A Fine-grained Approach for Power Consumption Analysis and Prediction*, Proceedings of the International Conference on Computational Science ICCS, Australia, 10-12 June, 2014.
5. Yuekun Wang, **Youssef Mesri**, *Learning by neural networks under constraints for simulation in fluid mechanics*. Computers and Fluids J. 2022.
6. Lianfa Wang, Yvan Fournier, Jean-François Wal, **Youssef Mesri**, *Fluid Flow characterization using Graph Neural Networks*, Proceeding of the Parallel CFD conference, 25-27 May, 2022.
7. R. Alomairy, W Bader, H. Ltaief, **Y. Mesri**, D Keyes, *High-performance 3D Unstructured Mesh Deformation Using Rank Structured Matrix Computations,* ACM Transactions on Parallel Computing 9 (1), 1-23, 2022.
8. **Youssef Mesri**, Hugues Digonnet, Thierry Coupez, *Advanced parallel computing in material forming with CIMLib*, European Journal of Computational Mechanics 18(7-8) pp. 669-694, 2009.
9. Nicholas Higham and T. Mary, *Mixed precision algorithms in numerical linear algebra*, hal-03537373, 2022
10. Rui Pereira, Marco Couto, Francisco Ribeiro, Rui Rua, Rui Jacome Cunha, Jao Paulo Fernandes, and Jao Saraiva, *Energy Efficiency across Programming Languages: How Do Energy, Time, and Memory Relate?*, Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering, New York, NY, USA, 2017.

Positionnement par rapport à l'axe de recherche TTI.5 retenu : (1/2 page maximum)

Ce projet s'insère dans l'axe de recherche **2 – *Une planète électrique*.**

L'empreinte carbone des systèmes électrique est celle de la part des sources fossiles dans la *production* et la *consommation* de l'électricité. Or les sources fossiles représentent pour l'instant une part importante du mix énergétique. Par exemple, le taux de *production de l'électricité renouvelable* proche des 100% n'est atteint en Europe qu'en Islande ou Norvège. Dans d'autres pays, la décarbonation de l'électricité n'est annoncée qu'au terme de plusieurs décennies.

En ce qui concerne l'IA, son empreinte carbone et son rôle dans le changement climatique a récemment fait l'objet d'un examen minutieux [1]. La prise de conscience a été déclenchée par le constat que l'entraînement d'un seul grand modèle de langage (NLP) pourrait s'approcher de 300 000 kg d'émissions de dioxyde de carbone [2], soit cinq fois les émissions à vie d'une voiture moyenne. En effet, l'IA, étant gourmande en calculs, est un important émetteur de carbone. Néanmoins, la solution n'est pas de prohiber l'IA, en effet, elle peut également jour un rôle positif et être utilisée pour réduire les effets du changement climatique en aidant à la conception de réseaux intelligents, en développant des infrastructures à faibles émissions et en modélisant les prévisions du changement climatique.

Il est donc crucial d'étudier comment réduire la consommation électrique et par conséquent l'empreinte carbone des calculs utilisés par les outils de l'IA. Ce sujet sera étudié dans le cadre de la présente thèse.

[1] P. Dhar, *The carbon impact of artificial intelligence*, Nature Machine Intelligence 2, 423–425 (2020). https://doi.org/10.1038/s42256-020-0219-9

[2] E. Strubell, A. Ganesh, and A. McCallum, *Energy and policy considerations for deep learning in NLP*, (2019) arXiv preprint arXiv:1906.02243.

Prismes d'analyse[1] adoptés par la thèse :

- Offre technologique,
- Opération et flexibilité des systèmes,
- Modes de vie et société,

[1] Les prismes d'analyse définis par TTI.5 sont : Offre technologique; Vecteurs et ressources; Opération et flexibilité des systèmes; Spatialité et rythme de mise en place; Aménagement du territoire; Modes de vie et société; Coûts et ingénierie de financement; Gouvernance; régulation et conditions institutionnelles; Externalités et impacts environnementaux.