LISTIC

Offre de stage 2024/2025

Titre	**Stratégies de distribution des données pour un apprentissage frugal et soucieux de l'empreinte carbone avec des sources d'énergie hétérogènes.**
Niveau	Master/Ingénieur
Date de début/ fin	Février / Juillet 2025
Ville, Pays	Annecy, *France*
Laboratoire	Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance - LISTIC
Description du sujet	De nos jours, l'intelligence artificielle révolutionne l'informatique et la vie quotidienne de chacun. Lors de l'utilisation des techniques d'apprentissage automatique, le volume important de données et de calculs nécessaires ne peut pas être pris en charge par un seul ordinateur. C'est pourquoi l'apprentissage automatique distribué (DML) est largement utilisé. Il existe de nombreuses façons de répartir efficacement la charge de travail. L'une de ces approches est l'informatique distribuée [Djebrouni2023], qui décompose les tâches en parties plus petites réparties sur plusieurs machines. L'apprentissage fédéré [Yang2019] est une autre stratégie prometteuse, dans laquelle l'apprentissage du modèle s'effectue localement sur des appareils ou des serveurs périphériques, ce qui réduit la nécessité d'un transfert de données important et permet d'économiser de l'énergie, tout en préservant la vie privée. Toutefois, ces stratégies sont à la fois gourmandes en données et en calculs, et également en termes de coûts de stockage ou de coûts réseau. La consommation d'énergie devient donc un problème essentiel, ce qui impacte directement leur empreinte carbone. L'objectif de ce stage est de répartir de manière efficace les données en fonction à la fois de la consommation énergétique des unités de calcul considérées mais surtout de l'empreinte carbone liée au mix énergétique de l'électricité consommée par les différentes unités de calcul. Certains auteurs ont commencé à étudier ce problème plus spécifiquement dans le cadre des datacenters (cf. [Vaconcelos2023], [Madon2020]). Dans ce cadre, des modèles de consommation énergétique ont également été proposé : [Heddeghem2012] présente un modèle pour l'estimation de l'empreinte carbone des datacenters distribués. En particulier, ils étudient la manière de distribuer les datacenters pour minimiser la consommation d'énergie et l'empreinte carbone. [Rostirolla2021] et [Grange2022] quant à eux étudient le cas des datacenter "vert", i.e., la conception de datacenter fonctionnant uniquement avec des énergies renouvelables locales. Dans ce stage, nous souhaitons étudier l'apprentissage fédéré avec des machines hétérogènes, distribuées sur des territoires plus ou moins éloignés avec des mix énergétiques différents qui évoluent en fonction du temps. Notre solution se veut la plus générale possible et étudiera donc des réseaux hétérogènes, que ce soit pour des apprentissages de type fédéré ou totalement distribué. Dans le cadre d'un précédent stage, un simulateur nommé FaLaFElS [falafels2024] a été développé. Il permet de calculer la consommation énergétique de l'apprentissage fédéré pour différentes topologies réseaux. L'objectif de ce projet est de faire évoluer ce simulateur pour prendre en compte l'impact carbone de la consommation énergétique des différentes unités de calcul du système distribué. Ce projet étudiera à la fois le cas off-line ou les sources énergétiques n'évoluent pas au cours du temps et le cas on-line avec des évolutions du mix énergétique au cours du temps. Pour ce faire, l'étudiant pourra s'appuyer sur les données de RTE pour les topologies et les empreintes carbones des sources d'énergie au niveau français (site web), mais également sur des topologies internationales (avec par exemple les données de electricitymaps). **Objectifs et contribution :** Ce stage peut être séparé en deux objectifs principaux : (O1) Mise en place d'une méthode de calcul de l'empreinte carbone d'un système distribué avec prise en compte de la position et de l'impact carbone des serveurs. (O2) Implémentation de politique de placement de données pour permettre de réduire l'empreinte carbone d'un système distribué. L'ensemble de ces deux objectifs a pour but d'être intégré dans l'outil FaLaFElS et testé pour différents types de systèmes distribués.

LISTIC

	Méthodologie : • Phase 1 : Réalisation d'une revue de littérature sur l'impact carbone et la consommation d'énergie des systèmes distribués, et des typologies des outils existants. • Phase 2 : Prise en main de l'outil de simulation FaLaFElS. • Phase 3 : Conception et implémentation d'une méthode de calcul d'empreinte carbone à partir des données d'entrée des différentes sources de données, avec une adaptation à l'outil de simulation FaLaFElS. • Phase 4 : Proposition d'algorithmes en ligne de répartition des données en tenant compte de la méthode de calcul de l'empreinte carbone établi en Phase 3 et en prenant compte de l'ensemble des contraintes du problème : contraintes géographiques / réseaux / latences. **Références :** [Djebrouni2023] Y. Djebrouni et al. Characterizing distributed machine learning workloads on apache spark: (experimentation and deployment paper). In Middleware ACM, 2023. [Yang2019] Q. Yang et al. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2):1–19, 2019. [Falafel2024] https://github.com/PhoqueEberlue/falafels [Savazzi2022] An Energy and Carbon Footprint Analysis of Distributed and Federated Learning; Stefano Savazzi, Vittorio Rampa, Sanaz Kianoush, and Mehdi Bennis. [Heddeghem2012] Distributed computing for carbon footprint reduction by exploiting low-footprint energy availability, Ward Van Heddeghem, Willem Vereecken, Didier Colle, Mario Pickavet, Piet Demeester. [Rais2019] Discover, model and combine energy leverages for large scale energy efficient infrastructures, Issam Raïs, Phd thesis [Vaconcelos2023] Optimal sizing of a globally distributed low carbon cloud federation; Miguel Vasconcelos, Daniel Cordeiro, Fanny Dufossé, Jean-Marc Nicod, Veronika Sonigo.[Madon2020] Integrating Pre-Cooling of Data Center operated with Renewable Energies; Maël Madon, Jean-Marc Pierson; IThings/GreenCom/CPSCom/SmartData/Cybermatics2020 [Rostirolla2021] A survey of challenges and solutions for the integration of renewable energy in datacenters; Gustavo Rostirolla, Léo Grange, Minh-Thuyen Thi, Patricia Stolf, Jean-Marc Pierson, Georges da Costa, Gwilherm Baudic, Marwa Haddad, Ayham Kassab, Jean-Marc Nicod, Laurent Philippe, Veronika Sonigo, Robin Roche, Berk Celik, Stéphane Caux, Jérôme Lecuivre. Renewable and Sustainable Energy Reviews, Elsevier, 2021, 155, pp.111787-1118 [Grange2022] Léo Grange, Patricia Stolf, Georges da Costa, Paul Renaud-Goud; Multi-Objective and Cooperative Power Planning for Datacenter With On-Site Renewable Energy Sources. IEEE Access, 2022. [Qiu2023] Xinchi Qiu, Titouan Parcollet, Javier Fernández-Marqués, Pedro P. B. de Gusmao, Yan Gao, Daniel J. Beutel, Taner Topal, Akhil Mathur, Nicholas D. Lane; A First Look into the Carbon Footprint of Federated Learning. J. Mach. Learn. Res. 24: 129:1-129:23 (2023)
Compétences requises	• Connaissance en système distribué, réseau, et en optimisation • Apprentissage distribué • C++
Gratification	Selon législation en vigueur (~600€/mois)
Tuteurs / Contacts	Plassart Stéphan, Monnet Sébastien (firstname.lastname@univ-smb.fr)