

Offre de stage 2026/2027

Titre	Stratégies d'optimisation et d'économie d'énergie pour l'entraînement de modèles d'IA sur des ressources de calcul hétérogènes
Niveau	Master/Ingénieur
Date de début/ fin	Septembre 2026 / Février 2027
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. Cependant, elle représente également une menace en termes de consommation énergétique et d'impacts écologiques indirects, tels que les émissions de carbone [Patterson2021], l'épuisement des ressources naturelles [Berthelot2024] et la consommation d'eau potable [Li2025]. Actuellement, l'approche dominante pour l'entraînement de modèles de pointe consiste à répartir la charge de travail sur de nombreuses unités de calcul avec des puissances homogènes. Bien que cette technique permet de maximiser le débit de calcul, cela implique de fortes contraintes matérielles et nécessite des investissements importants. Au contraire, effectuer les entraînements sur des clusters hétérogènes permettrait de réutiliser les infrastructures existantes, limitant les impacts indirects de l'IA. Dans ce contexte, de nouvelles problématiques émergent, tant scientifiques que logicielles. Par exemple, la synchronisation des paramètres (poids et biais) entre les différentes machines introduit des temps d'attente exacerbés par l'hétérogénéité des plateformes, représentant des pertes énergétiques évitables [MHDB2025]. De plus, l'hétérogénéité est également présente à l'échelle de la machine individuelle : Non-Uniform Memory Access (NUMA), GPU et/ou iGPU, Intel performance core / efficiency core etc. Utiliser ces ressources de manière optimale est donc un enjeu important qui pourrait permettre d'améliorer la frugalité des entraînements d'IA [Beamont2024]. Dans ce cadre, le framework DAHL est actuellement développé au sein du laboratoire pour répondre à ces problématiques. Il se base lui-même sur la librairie StarPU qui permet l'exécution et l'ordonnancement de tâches tant sur CPU que GPU [Thibault2018]. L'objectif de ce stage est d'explorer les différents leviers permettant de réduire la consommation électrique de l'entraînement d'IA. En plus des différentes méthodes liées à l'outil cité précédemment, il est également possible d'utiliser des méthodes de power capping [dePiolant2025] (Intel RAPL, Nvidia NVML) pour limiter la puissance des unités de calculs. Il serait intéressant d'étudier dans quelle mesure réduire la puissance permet d'économiser de l'énergie malgré les ralentissements que cela induit. Pareillement, il est possible de représenter les paramètres des modèles sur de plus petits nombres flottants, ajoutant une dimension de compromis entre la précision du modèle et l'utilisation de la mémoire/puissance de calcul. Dernièrement, StarPU permet d'utiliser un système d'ordonnancement basé sur des estimations de consommation. Cette approche novatrice semble intéressante appliquée dans ce contexte, et son impact reste peu étudié dans ce domaine. Ce cadre implique donc l'apprentissage et la manipulation de différents concepts et outils : premièrement il est nécessaire de comprendre la façon dont les calculs d'IA sont parallélisés. Dans notre cas, il s'agit d'un réseau neuronal à convolution (CNN) entraîné sur différents jeux de données classiques (Fashion-MNIST, CIFAR-10). Ensuite l'environnement de développement est basé sur un framework écrit en C et CUDA, ainsi que de nombreux outils permettant la recherche et le benchmarking (Valgrind, Alumet, Vite...). Dernièrement, l'acquisition des données via un wattmètre, le traitement des données, et la génération des graphiques est actuellement faite en Python. Objectifs et contribution : Ce stage peut être séparé en deux objectifs principaux : (O1) Mise en place et intégration des différentes méthodes d'économie d'énergie dans DAHL, (O2) Étude et comparaison des résultats, quels sont les compromis et limites des différentes solutions ?

	<u>Méthodologie :</u> - Phase 1 : Réalisation d'une revue de littérature sur les leviers de réduction de consommation énergétique et sur la parallélisation d'entraînement d'IA. - Phase 2 : Prise en main du framework DAHL et de la librairie StarPU. - Phase 3 : Mise en place, intégration et développement des différentes solutions d'économie d'énergie. - Phase 4 : Analyse et comparaison des solutions, identification des bottlenecks limitant les performances, propositions et itérations d'autres méthodes avantageuses. <u>Références :</u> [Li2025] Li, P. et al. Making ai less' thirsty'. Communications of the ACM, 68(7), 54-61. [Patterson2021] Patterson, D. et al. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350. [Berthelot2024]. Berthelot, A. et al. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Procedia CIRP, 122, 707-712. [dePiolant2025] de Piolant, A. et al. Improving energy efficiency of HPC applications using unbalanced GPU power capping. In 2025 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW) (pp. 820-829). IEEE. [Beaumont2024] Beaumont, O. et al. Exploiting Processor Heterogeneity to Improve Throughput and Reduce Latency for Deep Neural Network Inference. In 2024 IEEE 36th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (pp. 37-48). IEEE. [Thibault2018] Thibault, S. et al. On runtime systems for task-based programming on heterogeneous platforms (Doctoral dissertation, Université de Bordeaux). [MHDB] Mary Huet de Barochez, A. Plassart, S., Monnet, S. Falafels: A tool for Estimating Federated Learning Energy Consumption via Discrete Simulation. In International Conference on Advanced Information Networking and Applications (pp. 141-152). Cham: Springer Nature Switzerland. [StarPU2008] https://starpup.gitlabpages.inria.fr/ [Alumet2024] https://github.com/alumet-dev/alumet [DAHL2025] https://codeberg.org/PhoqueEberlue/dahl
Compétences requises	- Connaissances systèmes, aisance dans l'écosystème Linux - Intelligence Artificielle - C (requis), Python (utile)
Gratification	Selon législation en vigueur (~600€/mois)
Tuteurs / Contacts	Mary Huet de Barochez Andrew, Plassart Stéphan, Monnet Sébastien (firstname.lastname@univ-smb.fr)