Optimisation du parallélisme de pipeline pour les grands modèles de langage (LLMs) via l'exploitation du parallélisme interne des opérations

Encadrement : Olivier Beaumont, Lionel Eyraud-Dubois, Julia Gusak (Topal, Centre Inria de l'Université de Bordeaux)

Ce sujet s'inscrit dans le cadre d'un projet de recherche en cours de montage, avec une forte probabilité de financement pour une poursuite en thèse. Le projet global concerne la conception de nouveaux accélérateurs dédiés à l'IA, avec un focus particulier sur la parallélisation de l'apprentissage et de l'inférence.

Le travail proposé porte sur l'optimisation du parallélisme de pipeline dans l'exécution de modèles d'IA. On considère un modèle représenté sous la forme d'un graphe orienté acyclique (DAG), où chaque nœud correspond à une opération du modèle et est annoté par ses temps d'exécution selon le nombre d'accélérateurs utilisés (1, 2, …, p). Dans un premier temps, les coûts de communication seront négligés, mais pourront être réintroduits dans une seconde phase du travail.

Le traitement d'un micro-batch correspond à une traversée complète de ce graphe. L'objectif est d'exécuter ce DAG en mode pipeline, en injectant de nouveaux micro-batches aussi fréquemment que possible, afin de maximiser le débit. Si ce problème a été abondamment étudié dans le cas simple de graphes linéaires composés de tâches de coût unitaire, le cas de dépendances plus complexes et de tâches hétérogènes reste largement ouvert.

Pourtant, il est crucial pour réduire la granularité des tâches et exploiter plus finement le parallélisme intrinsèque des graphes, en particulier dans le contexte des LLMs.
Dans un premier temps, le travail portera sur des graphes correspondant à des modèles de type LLM, constitués d'une séquence de blocs Transformer, chaque bloc étant lui-même modélisé comme un petit DAG. À plus long terme (notamment dans le cadre de la thèse), ce DAG sera généré automatiquement par une chaîne de compilation dédiée aux nouveaux accélérateurs développés dans le projet.

Le sujet est volontairement ouvert et modulable. Il peut inclure une part plus ou moins importante de travail algorithmique, théorique et/ou d'implémentation, selon le profil et les intérêts du candidat. Il s'agit d'un sujet à l'interface entre algorithmique, HPC et IA : une appétence pour l'algorithmique et les architectures parallèles est essentielle, et un intérêt pour l'IA, même sans expertise préalable approfondie, est bienvenu.