Internalisation de ressources IA : pour concilier sobriété énergétique et souveraineté technologique

Emmanuel Quemener*[1]

[1]École normale supérieure de Lyon – Université de Lyon, Centre Blaise Pascal en Modélisation et Sciences Numériques – France

Résumé

Il ne se passe pas un jour sans que l'IA et ses énergivores mercenaires, les Data Centers, soient accusés des pires maux en matière de réchauffement climatique, de consommation hydrique ou d'épuisement des ressources : mais sans en être les principaux responsables, ils le sont à juste titre...
Dans un contexte géopolitique tendu (suivant la crise énergétique) et des budgets faméliques, comment satisfaire des besoins croissants en IA tout en contrôlant son empreinte écologique et sa dépendance à des infrastructures dont les coûts financiers ou enironnementaux nous sont inconnus ?

L'internalisation complète de ces outils offre, presque paradoxalement, des solutions à ces défis. Et nous nous pencherons sur des réalisations en la matière du CBPsmn (Centre Blaise Pascal en Modélisation et Sciences Numériques).

Tout d'abord, nous nous pencherons sur l'empreinte environnementale des équipements. Avec une électricité très largement décarbonée, la France peut (presque) réduire son empreinte environnementale IT à la fabrication des équipements, dès lors qu'elle en internalise les équipements... Ainsi les exploiter le plus longtemps possible constitue une approche vertueuse. Mais est-il pertinent d'exploiter des machines anciennes en IA ? Pour quelles tâches, de l'inférence ou de l'affinage ? Nous verrons dans quels cas d'usages cela reste intéressant, à condition cependant de choisir sa GPU de manière pertinente.

Puis, nous aborderons la nature des traitements IA qu'il est possible d'intégrer "à moindre frais" humain dans les infrastructures. Curieusement, le @TheEdge, le déploiement sur la périphérie hors du Data Center, offre des perspectives intéressantes : il permet l'exploitation de machines avec des composants "génériques", largement disponibles, mais exigeant un changement de paradigme dans le "placement" des équipements de calcul haute performance.

Ensuite, nous nous intéresserons aux solutions destinées à exploiter ce "déchet permanent" de l'équipement IT : la chaleur fatale. Cela, en analysant puis en déployant des solutions optimisant la consommation voire en exploitant complètement cette chaleur : par exemple les AnchIAles, ces auxiliaires de chauffage dopées à la carte graphique de Gamer parfaites pour de l'inférence, mais cependant moins efficaces pour de l'entraînement.

Enfin, nous clôturerons que ces solutions matérielles ne sont pas parfaites ; elles ont toutefois

*Intervenant

le mérite de limiter la surface de dépendances aux infrastructures externes, notamment en proposant en accès simplifié des modèles de LLM complètement internalisés.
Nous conclurons que dans un contexte de pénurie (relative) de matériel neuf et face à l'émergence de ces nouveaux besoins IA, il est toujours possible d'offrir des solutions à coûts maîtrisés : seulement du matériel. Contrairement aux idées reçues, la marche vers cette autonomie (relative) de son environnement IT n'est pas plus haute que la constellation de toutes les certifications sur des matériels et leur exploitation souvent bancale.