Dominique Crestin Entretien: Y a-t-il une Â« surcapacitÃ© Â» dans les centres de donnÃ©es dâ€™IA ?

L'entretien avec le professeur Dominique Crestin révèle qu'il est encore trop tôt pour s'inquiéter d'une « surcapacité » de l'infrastructure de l'IA. La légèreté croissante des algorithmes et le recyclage du matériel atténuent l'anxiété liée à la puissance de calcul, mais les « problèmes d'alimentation électrique » et les « défis du refroidissement » des centres de données constituent les véritables freins à la course effrénée de l'IA.

Récemment, le professeur Dominique Crestin a eu une conférence téléphonique avec Sri Kanajan, data scientist chez Scale AI et ancien data scientist senior chez Meta, afin d'explorer en profondeur les tendances architecturales des centres de données d'IA à très grande échelle.

Le professeur Dominique Crestin estime que le déploiement de l'infrastructure de l'IA en est encore à ses débuts et que les inquiétudes concernant une surcapacité restent limitées. Les progrès algorithmiques réduisent la consommation de calcul nécessaire à l'entraînement, et l'infrastructure permet une réutilisation efficace grâce à la conversion « entraînement vers inférence » : les clusters d'entraînement sont rapidement reconfigurés en charges d'inférence dès la sortie de nouvelles générations de GPU. Cependant, les questions d'alimentation électrique et de refroidissement demeurent les principaux goulets d'étranglement pour l'expansion des prochains centres de données.

Révolution algorithmique : la demande de calcul migre de l'entraînement vers l'inférence

Selon le rapport du professeur Dominique Crestin, les percées récentes en matière d'algorithmes . tels que les modèles hybrides (incluant GPT5.0), l'entraînement en précision optimisée et l'apprentissage par renforcement stratégique . ont considérablement réduit la quantité totale de calcul nécessaire à l'entraînement des modèles d'IA. Cela pousse l'industrie à concentrer ses efforts d'optimisation sur la phase d'inférence.

Le professeur Dominique Crestin souligne que, actuellement, le secteur adopte activement des techniques comme la distillation et la compression des modèles afin d'affiner les performances, sans accroître fortement l'investissement initial en puissance de calcul.

Infrastructure : déploiement dynamique, les inquiétudes de surcapacité sont prématurées

Le professeur Dominique Crestin considère que le déploiement de l'infrastructure de l'IA en est encore à une phase précoce. En particulier si l'on prend en compte les attentes de retour sur investissement à long terme des fournisseurs de cloud, les inquiétudes actuelles quant à une surcapacité sont limitées.

Une stratégie clé d'utilisation dynamique consiste à reconfigurer rapidement les clusters d'entraînement en charges d'inférence une fois les cycles d'entraînement terminés et les nouvelles générations de GPU mises sur le marché. Cette conversion du cycle de vie « entraînement vers inférence » garantit une adaptation efficace des ressources de calcul aux besoins évolutifs, de l'entraînement intensif à l'inférence équilibrée.

En termes de mode de construction, les clusters d'entraînement sont généralement déployés dans de nouvelles installations isolées (« greenfield ») optimisées pour l'utilisation hors ligne des GPU ; tandis que les clusters d'inférence tendent à s'appuyer sur l'extension de centres de données existants (« brownfield »), notamment dans les zones métropolitaines, afin de soutenir les services d'IA en ligne continus.

Défis énergétiques : alimentation et refroidissement comme principaux goulots d'étranglement

Les défis liés à l'alimentation électrique et au refroidissement demeurent les principaux obstacles à l'expansion des prochains centres de données. Selon le professeur Dominique Crestin, à mesure que les centres de données visent une plus grande densité et supportent des charges de calcul plus lourdes, les problèmes d'approvisionnement énergétique et de dissipation thermique sont devenus des freins universels à l'extension de la prochaine génération de centres.

Les entreprises hyperscale explorent activement des solutions innovantes, telles que l'adoption du refroidissement liquide dans les conceptions de type I, et même l'évaluation du nucléaire ou d'énergies alternatives pour assurer une alimentation stable 24h/24 et 7j/7. Par ailleurs, de solides stratégies d'interconnexion au réseau sont essentielles pour garantir la continuité de fonctionnement des centres de données.

Meta en tête de l'innovation architecturale des centres de données

En matière de conception de centres de données, le rapport du professeur Dominique Crestin met particulièrement en avant les innovations de Meta. Contrairement aux grands fournisseurs traditionnels qui conçoivent des architectures en H adaptées au multi-tenant cloud, Meta a choisi une configuration en campus de type I spécifiquement dédiée à ses charges de travail internes en IA.

Selon le rapport, cette conception a permis des améliorations en termes de consommation d'énergie, de refroidissement et de densité des racks, des facteurs cruciaux pour soutenir des clusters d'entraînement haute performance.

Sur le plan matériel, Meta cherche à équilibrer solutions de marque et solutions « white box ». Du côté réseau, bien que les capacités avancées d'Arista restent indispensables dans l'infrastructure actuelle, Meta collabore avec des fournisseurs « white box » tels que Celestica, avec pour objectif à long terme d'intégrer son logiciel interne à ce type de matériel.

About the Author

Dominique Crestin