La puissance phénoménale du nouveau monstre de l’IA : ChatGPT en mode turbo grâce à son GPU surpuissant !

Nvidia dévoile le GPU H200 : une révolution pour l’IA

Nvidia dévoile le GPU H200 : une révolution pour l’IA

Après avoir conquis tout l’écosystème IA avec le GPU H100, Nvidia dévoile le H200 : la nouvelle référence des cartes graphiques dédiées à l’intelligence artificielle. Découvrez comment cette bombe de puissance va permettre aux IA existantes comme ChatGPT de passer au prochain niveau, tout en donnant naissance à de nouveaux modèles encore plus avancés !

Jadis, Nvidia était principalement connue par les joueurs de jeux vidéo sur PC pour ses puissantes cartes graphiques. Toutefois, au cours des dernières années, l’entreprise américaine s’est hissée parmi les maîtres du monde. Avec la mode des cryptomonnaies, d’abord, ses GPU ont été massivement utilisés pour le « minage » de Bitcoin et autres. Plus récemment, avec l’essor de l’IA générative, ses puces servent désormais pour l’entraînement des modèles comme GPT. Et c’est précisément ce qui a permis à Nvidia de rejoindre le club très fermé des entreprises capitalisées à plus d’un trillion de

dollars, car l’entraînement de l’IA requiert beaucoup, beaucoup de cartes graphiques.

Les GPU sont idéaux pour les applications IA, car ils peuvent effectuer de nombreuses multiplications de matrices parallèles, indispensables au fonctionnement des réseaux de neurones. Leur rôle est essentiel pour l’entraînement et l’inférence des modèles IA. Comme l’explique Ian Buck, vice président du HPC chez Nvidia, « pour créer des l’intelligence avec de l’IA générative et des applications HPC, de vastes volumes de données doivent être traités efficacement à haute vitesse en utilisant la mémoire large et rapide des GPU ».

À présent, les GAFAM et tous les acteurs de l’industrie de l’IA comme OpenAI et Anthropic se livrent une guerre sans merci pour accaparer autant de GPU que possible. À tel point qu’une pénurie se profile pour le grand public.

Le nouveau champion des cartes graphiques IA

De son nom complet HGX H200 Tensor Core, ce nouveau GPU utilise l’architecture Hopper pour accélérer les applications IA. Et il pourrait permettre la création de modèles IA encore plus puissants. En outre, grâce à ce composant, les IA déjà existantes comme ChatGPT pourraient profiter d’un temps de réponse fortement amélioré. Selon Nvidia, le H200 est tout simplement le premier GPU à proposer de la mémoire HBM3e. Cela lui permet de délivrer 141GB de mémoire et 4,8 terabytes par seconde de bande passante. À titre de comparaison, cela représente 2,4 fois la bande passante du Nvidia A100 lancé en 2020. Autant dire que la technologie évolue vite, très vite.

La puissance comme remède à la pénurie ?

Tous les experts s’accordent à dire que le manque de puissance a été l’un des principaux freins au progrès de l’IA au cours

de l’année 2023. Cette lacune a ralenti le déploiement des modèles IA existants et le développement des nouveaux. Or, la pénurie de puissants GPU IA est l’une des principales causes. Par exemple, OpenAI a souvent répété que le manque de GPU ralentit ChatGPT. La firme est contrainte de limiter son chatbot pour pouvoir délivrer le service. L’une des solutions les plus évidentes est donc de produire davantage de puces, mais créer des puces plus puissantes contribue également à résoudre le problème. Grâce au H200, les modèles IA sur lesquels repose ChatGPT pourront servir davantage de clients simultanément.

Prix et date de lancement

Le H200 sera disponible dans plusieurs formats. Le HGX H200 est une carte serveur proposée en configurations four-way ou eight-way, compatible avec le hardware et le logiciel des systèmes HGX H100. De son côté, la Nvidia GH200 Grace Hopper Superchip combine un CPU et GPU en un seul package pour encore plus de puissance IA. Ces différentes options répondront à tous les besoins. Les premières instances basées H200 seront déployées à partir de 2024 par Amazon Web Services, Google Cloud, Microsoft Azure et Oracle Cloud Infrastructure. Par la suite, le GPU sera disponible auprès de tous les fournisseurs de services cloud et les fabricants de systèmes à partir du deuxième trimestre 2024. Le prix n’a pas encore été communiqué.