ChatGPT propulsé par le géant Nvidia : la nouvelle carte graphique révolutionnaire aux capacités impressionnantes !

Plateforme Nvidia GH200 Grace Hopper : un nouveau processeur pour l'IA générative

Lors du salon SIGGRAPH, Nvidia présente sa nouvelle plateforme pour l'IA générative

Au salon SIGGRAPH consacré à l'imagerie par ordinateur, Nvidia a dévoilé la dernière génération de sa plateforme pour l'IA générative et les superordinateurs. Voici la plateforme Nvidia GH200 Grace Hopper, la première carte graphique HBM3e au monde.

Source : Nvidia / Computex

Nvidia est principalement connu pour ses cartes graphiques destinées aux jeux vidéo, les célèbres GeForce GTX et RTX. Cependant, la société est également présente dans le domaine

professionnel grâce à la puissance de calcul brute de ses produits. Grâce à la popularité croissante de l'IA générative grâce à des outils tels que ChatGPT ou Midjourney, Nvidia a connu un grand succès, en vendant des puces utilisées par les géants de l'IA pour leurs serveurs de calcul. La valeur des actions de Nvidia a augmenté de quatre fois entre octobre 2022 et août 2023, et l'entreprise est maintenant évaluée à plus de 1100 milliards de dollars.

Grace Hopper GH200 : la mémoire n'est plus un problème

Nvidia présente cette semaine une nouvelle solution pour ce marché avec la plateforme Grace Hopper Nvidia GH200. Comme son nom l'indique, cette plateforme combine une carte graphique Hopper avec un processeur Nvidia Grace. Cette nouvelle génération met surtout l'accent sur la bande passante et la capacité mémoire. Du côté du processeur, Grace intègre 144 cœurs ARM Neoverse. La puissance de calcul pour l'IA est annoncée à 8000 téraflops.

Source : Nvidia

Nvidia promet une bande passante multipliée par 3, atteignant 10 To/s, et une capacité mémoire pouvant aller jusqu'à 282 Go de mémoire HBM3e à très haute vitesse. On est loin des controverses autour des 8 Go de VRAM sur les cartes graphiques grand public.

Cette nouvelle plateforme sera intégrée dans des serveurs à partir du printemps 2024. L'augmentation de la bande passante permettra une meilleure mise à l'échelle de ces puces, augmentant ainsi la puissance de calcul générale de l'infrastructure. En d'autres termes, les outils les plus gourmands, notamment dans le domaine de l'IA, pourront bénéficier de calculs plus performants. Pour les utilisateurs,

cela se traduira par une génération d’images plus rapide, par exemple.

Pour nous suivre, nous vous invitons à télécharger notre application Android et iOS. Vous pourrez y lire nos articles, dossiers et regarder nos dernières vidéos YouTube.