

S'ABONNER

Économie Technologie Marketing Intelligence artificielle Retail Cybersécurité IT Médias Réseaux sociaux

Intelligence artificielle

Pourquoi Nvidia a toujours eu « un coup d'avance sur ses concurrents » ?

Initialement conçus pour le jeu vidéo, les GPU de Nvidia sont désormais exploités pour le développement et le fonctionnement d'IA génératives.

f t in e



Photographie : Nvidia.

Par **Zacharie Tazrout** - @Zach_Tzt

Publié le 20 mars 2024 à 08h18

Nvidia s'est imposé comme l'entreprise incontournable de l'intelligence artificielle (IA). Les grands groupes comme les start-up s'arrachent à prix d'or ses composants, capables de former des modèles d'IA toujours plus gourmands en ressources. Pour quelles raisons les semi-conducteurs de la firme de Jensen Huang sont-ils aussi prisés, par rapport à ceux de ses concurrents ?

Avant d'être le leader des puces IA, Nvidia a été le premier à s'intéresser au calcul haute performance

Fin 2022, **OpenAI dévoile ChatGPT**, un robot conversationnel marquant le début de l'engouement autour de l'IA générative. La start-up dirigée par Sam Altman est rapidement imitée. D'abord par les géants technologiques, et ensuite par plusieurs start-up prometteuses comme **Anthropic** ou **Mistral AI**. Toutes forment de grands modèles de langage (LLM) performants, capables d'alimenter des IA génératives de type ChatGPT. Ces LLM sont composés de multiples couches d'opérations, nécessitant **une immense puissance de calcul**. Pour les former, ces entreprises se ruent sur les mêmes composants, les GPU développées par Nvidia.

« *Au milieu des années 2000, plusieurs chercheurs se sont intéressés au potentiel des unités de traitement graphiques pour le calcul haute performance. Tous travaillaient en collaboration avec Nvidia* », indique Didier El Baz, enseignant-chercheur au Laboratoire d'analyse et d'architecture des systèmes (LAAS) du CNRS, à Toulouse. Connue pour ses produits à destination du grand public, surtout des amateurs de jeux vidéo, l'entreprise voit un potentiel dans le développement de GPU à haute puissance de calcul, taillés pour le marché professionnel.

Inscrivez-vous à la newsletter

Votre adresse email

S'INSCRIRE

En vous inscrivant vous acceptez notre [politique de protection des données personnelles](#).

Rapidement, ils se rendent compte que ses puces sont nettement plus efficaces que les processeurs (CPU) pour traiter une grande quantité de données. Contrairement aux CPU, les GPU peuvent contenir plus de transistors et plus de cœurs.

Après avoir testé le potentiel des GPU pendant près de deux ans, les chercheurs développent une technologie qui va révolutionner leur usage : le General-Purpose Computing on Graphics Processing Units (GP-GPU). Elle permet d'utiliser des unités de traitement graphiques afin d'exécuter des calculs généraux à la place du CPU. Nvidia propose dès 2007 son propre GP-GPU, CUDA (pour Compute Unified Device Architecture), qui permet, entre autres, de mener des calculs parallèles.

« *Progressivement, ce “framework” est devenu un standard dans le domaine* », explique Yannis Duffourd, bioinformaticien à l'Université de Dijon et utilisateur régulier des composants Nvidia dans le cadre de ses travaux. Les

principaux concurrents de CUDA, OpenCL développé par Khronos Group sur une idée d'Apple, et ATI Stream d'AMD ne sortent que deux années plus tard, laissant le temps à Nvidia de peaufiner sa création.

Avec CUDA et le développement de la maîtrise des GPU dans la recherche, Nvidia détient dès la fin des années 2000, toutes les clés pour conquérir de nouveaux marchés. Il ne lui manque plus qu'une seule chose : étoffer ses capacités de production. Nvidia est une « fabless », elle design ses puces, mais a besoin d'un sous-traitant pour les fabriquer. Elle se tourne vers le taïwanais TSMC, à la fois le fabricant sous contrat le plus important du monde et celui le plus à la pointe de la technologie.

Nvidia a pris une avance considérable dans la course aux GPU

Les générations de processeur se succèdent, et Nvidia a toujours « *un coup d'avance sur ses principaux concurrents* » selon Didier El Baz. Le groupe ne se contente pas de proposer des GPU de haute volée, Il a « *participé au développement de l'écosystème autour des GPU afin de les démocratiser, faciliter leur utilisation, sans que les entreprises n'aient besoin de comprendre leur fonctionnement pour les exploiter* ». Au niveau logiciel, CUDA en est le meilleur exemple.

Pour ce qui est du matériel, Nvidia va plus loin en proposant aux entreprises ses propres supercalculateurs clé en main. En parallèle, l'entreprise se lance dans la conception de cœurs spécifiques, « des cœurs Tensor, une spécialité de l'entreprise qui existe nulle part ailleurs », précise le chercheur du LAAS-CNRS. « *Ces cœurs GPU sont beaucoup plus spécialisés et nécessitent une adaptation très importante du code logiciel pour être utilisés à leur plein potentiel, mais au final sur des tâches de calcul, vous êtes largement gagnants en termes de puissance, rapidité, etc.* », ajoute Yannis Duffourd.

À la fin des années 2010, Intel s'est intéressé de près au marché des GPU dédiés au calcul haute performance et à la formation d'IA. Toutefois, elle s'est rendu compte qu'elle accusait un retard technologique considérable sur Nvidia. Pour tenter de le combler, l'entreprise a racheté **Habana Labs**, une société israélienne spécialisée dans les accélérateurs d'IA. Ensemble, elles ont développé et commercialisé Gaudi en 2020. Malgré l'expertise d'Habana Labs, Gaudi ne possédait pas autant de transistors, ni des cœurs tensor que les GPU proposés par le groupe de Jensen Huang à cette période.

Avec Gaudi 2, sortie en 2022, Intel proposait enfin un GPU **avec des performances égales** à celles proposées par Nvidia en 2020. La prochaine génération, Gaudi 3, doit rivaliser avec les plus puissantes du marché actuellement disponibles. Malheureusement pour Intel, Nvidia continue également d'améliorer ses produits. Les H200 prévues pour la fin du semestre présentent des caractéristiques jamais vues, et les géants de la tech ont déjà assuré qu'ils les exploiteraient dans leurs infrastructures.

Le succès des produits de Nvidia est le fruit **de son histoire**. C'est au cours des années 2000, lorsque l'entreprise a décidé de s'intéresser au calcul haute performance que le premier jalon de sa réussite a été posé. On s'arrache ses unités de traitement graphiques pour la formation et le fonctionnement des grands modèles de langage parce que « *Nvidia a toujours travaillé pour proposer un composant qui soit vu comme le couteau suisse du calcul haute performance* » analyse à posteriori Didier El Baz. « *Aujourd'hui, ils y sont parvenus* » conclut-il.

Cet article vous est offert par Siècle Digital dans le but de vous faire découvrir ses formules d'abonnement. Convaincu(e) ? **Abonnez-vous.**

Pour lire la suite...

Inscrivez-vous à la matinale de Siècle Digital

Votre adresse email

☐ J'accepte que mon email soit collecté pour que Siècle Digital m'envoie des newsletters (requis)

Où vont mes données ?

Merci, j'en profite !

[Je me connecte](#)

[Non, merci](#)

Pour aller plus loin

Santé, jumeau numérique, quantique, IA... Nvidia fait le plein de nouveautés au GTC 2024

Progressivement, Intel rattrape son retard sur Nvidia en matière d'accélérateurs d'IA

Nvidia, d'une start-up gaming en faillite à géant de l'IA

Les géants de l'IA générative s'intéressent aux robots humanoïdes

Le blocage de la fusion Nvidia-Arm a stimulé l'innovation, estime Lina Khan

Sam Altman, les dessous d'un projet à 7 000 milliards de dollars

INTEL

NVIDIA

SEMI-CONDUCTEURS

Pratique

+

Ressources

+

Abonnement

+

Tous droits réservés - Siècle Digital 2023