TECH

Le CES, théâtre du bras de fer entre Nvidia et AMD pour les puces IA

LES DEUX GÉANTS SONT MONTÉS SUR SCÈNE POUR FAIRE UNE DÉMONSTRATION DE FORCE. AMD VEUT PROUVER QU'IL EST EN MESURE DE RIVALISER AVEC LE LEADER DU SECTEUR.

5 min • Lucas Mediavilla

AMD peut-il contester l'hégémonie de Nvidia dans les puces dédiées à l'IA ? Ce lundi, du côté du CES de Las Vegas, tous les regards étaient tournés vers les deux champions californiens et leurs PDG respectifs, Lisa Su et Jensen Huang (par ailleurs cousins germains), qui tenaient leur conférence dans le Nevada à cinq heures d'intervalle. Depuis l'essor de l'IA générative, Nvidia règne sans partage sur les processeurs permettant de faire fonctionner l'IA dans les data centers. Sa part de marché oscille entre 80 % et 90 % selon les analystes.

Mais AMD, historiquement positionné sur les puces pour PC, lui donne depuis quelques mois du fil à retordre. L'accord signé début octobre avec OpenAI, pour construire une infrastructure géante de data centers d'IA embarquant les puces d'AMD, a été perçu par les investisseurs et analystes comme une preuve de la montée en puissance de l'entreprise de Lisa Su. Ce dernier pourrait générer près de 100 milliards de dollars de revenus pour AMD. Forte de ce constat, la dirigeante s'est échinée à montrer que son entreprise était plus que jamais une alternative à son rival californien.

Très attendue, la dirigeante a notamment annoncé le lancement d'une nouvelle puce à GPU super puissante et dédiée aux usages les plus gourmands de l'IA - la MI455X. Cette dernière sera vendue dès 2026 aux entreprises du secteur, mais constitue surtout la pierre angulaire de l'accord signé avec OpenAI il y a quelques semaines. En 2027, une autre série, la MI500, pourrait offrir des performances jusqu'à 1 000 fois plus importantes que la série MI300 lancée en 2023.

Dans une stratégie visant à couvrir tous les besoins du marché, y compris ceux des plus petites entreprises, AMD a annoncé également le lancement de la puce MI440X. L'idée étant de pouvoir déployer ce type de composants en local sur les centres de données des entreprises, et ainsi ne pas avoir à transférer de données dans le cloud.

À l'image du patron de Nvidia, Jensen Huang, Lisa Su est convaincue que l'appétit des entreprises d'IA pour la puissance de calcul n'est pas près de se tarir. *« Nous n'avons pas assez de puissance de calcul pour tout ce que nous pourrions faire »*, a expliqué la dirigeante sur scène, indiquant qu'il faudrait d'ici cinq ans multiplier la puissance de calcul par 10 000 par rapport à 2022 pour répondre aux besoins de l'industrie. Une estimation qui ne va d'ailleurs pas dissiper les préoccupations de plus en plus fortes sur la consommation énergétique des centres de données.

Rejointe sur scène par Greg Brockman, le président d'OpenAI, mais également par les dirigeants d'entreprises comme Luma AI, Blue Origin, AstraZeneca ou Generative Bionics, Lisa Su a ainsi tenté de faire valoir la richesse de son écosystème de partenaires dans l'IA. Celui-ci s'étendant ainsi du champion incontesté des modèles de langage, OpenAI, à d'autres leaders dans des secteurs aussi variés que le médical, le spatial ou même les robots humanoïdes (Generative Bionics).

Une démonstration de force suffisante pour faire tressaillir Nvidia ? Rien n'est moins sûr. Le géant de San Jose n'est pas venu dans le Nevada les mains vides, même si les annonces de produits étaient moins nombreuses. Jensen Huang a notamment indiqué que les processeurs à GPU Rubin, annoncés en 2024, et censés remplacer l'architecture Blackwell qui s'arrache encore à ce jour chez tous les géants de l'IA, sont entrés en *« pleine production »*. Rubin promet des gains d'efficacité d'un -facteur dix par rapport à la précédente architecture. Le processeur sera livré et vendu à compter du second semestre de l'année en cours et comptera parmi ses premiers clients Microsoft ou encore CoreWeave. La demande *« est très forte »*, n'a pas manqué de rassurer Jensen Huang.

Sachant que la performance brute de ses puces reste au-dessus de celle de ses rivaux mais qu'il va nécessairement perdre de la part de marché, le PDG de Nvidia ne s'est pas arc-bouté sur la puissance de ses composants. Alors que ses gros clients sont aujourd'hui les fournisseurs de cloud (Oracle, Meta, AWS, Microsoft Azure), une majeure partie de sa conférence avait ainsi pour but de pousser encore un peu plus l'adoption de l'IA dans les autres secteurs de l'industrie.

Le principal relais de croissance identifié concerne l'IA dite physique, autrement dit la robotique (industrielle et humanoïde). Flanqué comme d'habitude à ses côtés de petits robots, Jensen Huang a ainsi égrené l'ensemble des initiatives de Nvidia dans le secteur de la robotique, depuis les puces qui permettent au robot de bouger et d'agir à l'infrastructure logicielle qui permettent d'entraîner le robot et le rendre - intelligent. Cet entraînement se faisant d'ailleurs au travers des plateformes de simulation virtuelle du groupe (Cosmos, notamment).

Au-delà de la robotique, c'est du côté de l'automobile que Nvidia avait réservé plusieurs annonces. « *Notre vision est qu'un jour, chaque voiture, chaque camion sera autonome* », a indiqué Jensen Huang ce lundi à Las Vegas. Dans le Nevada, le groupe a ainsi annoncé le lancement d'Alpamayo, considéré comme le premier modèle de raisonnement pour véhicule autonome.

En d'autres termes, il s'agit d'un ensemble de modèles d'IA, d'outils de simulation et de jeux de données qui doit permettre aux développeurs d'améliorer les systèmes de conduites autonomes. *« Alpamayo apporte des capacités de raisonnement aux véhicules autonomes, leur permettant d'analyser des scénarios rares, de circuler en toute sécurité dans des environnements complexes et d'expliquer leurs décisions de conduite »*, a promis sur scène Jensen Huang. Partenaire de Mercedes sur Alpamayo, Nvidia a annoncé des tests avec passagers dès le deuxième trimestre aux États-Unis. L.M.