# D'OpenAI à Google, les géants de l'IA lancent leurs propres puces

Alexandre Piquard

## Des acteurs de l'IA et du cloud veulent réduire leur dépendance à Nvidia

OpenAI espère, à partir de 2026, disposer de sa première puce maison. L'information, rapportée par le *Financial Times* le 5 septembre, n'a pas fait les gros titres mais elle illustre une tendance. La start-up d'intelligence artificielle (IA) s'est lancée dans la conception d'un processeur destiné à réaliser une partie des énormes calculs informatiques nécessaires pour répondre aux requêtes des utilisateurs de son célèbre assistant ChatGPT et pour entraîner ses futurs modèles d'IA. OpenAI a pour cela constitué une équipe de 40 personnes et s'est appuyé sur un partenaire, Broadcom. Ce spécialiste des semi-conducteurs a vu son titre bondir en Bourse de 9 % à cette annonce, que l'entreprise de Sam Altman ne commente pas officiellement.

Pourquoi une entreprise de logiciels cherche-t-elle à faire produire ses propres puces ? Le cas d'OpenAI est loin d'être isolé : la plupart des géants de la tech actifs dans l'IA ont commencé à concevoir des processeurs. *« Depuis le début de l'essor de l'IA, les grands acteurs achètent des puces à Nvidia. Mais ils se disent qu'il doit y avoir une alternative à cette situation où ils versent des milliards à cette entreprise aux marges très conséquentes »*, dit Sébastien Sztabowicz, analyste chez Kepler Cheuvreux. Nvidia, leader incontesté des processeurs dévolus à l'IA (appelés « GPU »), affiche plus de 70 % de marge brute. Et l'entreprise est devenue en juillet la première à passer la barre des 4 000 milliards de dollars (3 380 milliards d'euros) de capitalisation boursière…

A lui seul, l'achat de ses puces devrait représenter 28 % des 350 milliards de dollars que Google, Amazon, Meta, Microsoft et Oracle comptent dépenser en 2025, a calculé en juillet Goldman Sachs. Pour réduire leur *« dépendance »* à Nvidia, dit M. Sztabowicz, ces géants de l'IA et de l'hébergement dans le cloud conçoivent des puces, souvent spécialisées pour certains usages, et ils s'appuient sur des partenaires du secteur des semi-conducteurs comme Broadcom, MediaTek ou Marvell.

## Nvidia plutôt serein

Chez Amazon, on explique voir un intérêt économique dans la conception de puces et avoir déjà un savoir-faire. Le géant du cloud collabore avec Annapurna Labs, une start-up israélienne rachetée en 2015, avec laquelle il a déjà fait produire un processeur voué aux serveurs classiques, baptisé Graviton. Pour l'IA, Amazon a lancé en 2023 les puces Inferentia, consacrées aux calculs liés au fonctionnement des modèles (ou « inférence ») et les Trainium, dévolues à leur entraînement intensif initial.

L'entreprise assure que ses puces ont un rapport prix-performance de 30 % à 40 % meilleur que les puces Nvidia. Et dit avoir déjà comme utilisateurs SAP, Databricks, Poolside ou Anthropic, la start-up concurrente d'OpenAI. Avec cette dernière, Amazon compte passer à une autre échelle : elle a annoncé fin 2024 le projet Ranier, un data center géant de *« centaines de milliers »* de puces Trainium 2, pour un coût de 8 milliards de dollars. Google est bien avancé : le groupe a annoncé au printemps la septième génération de ses puces TPU. Celles-ci font fonctionner son modèle d'IA de pointe Gemini 2.5 ou l'application de biologie AlphaFold couronnée du prix Nobel.

Ses puces sont aussi utilisées via sa plate-forme cloud par Anthropic, Midjourney, Hugging Face ou le français Mistral AI, ajoute Google. Meta en est aux tests pour sa première puce capable d'entraîner ses IA, attendue en 2026, a rapporté Reuters en mars. Elle dispose déjà de processeurs baptisés « MTIA », mais ils sont cantonnés au fonctionnement d'IA comme les algorithmes de classement des contenus sur Facebook ou Instagram, ajoute l'agence.

Les initiatives des groupes de tech dans cette nouvelle activité rencontrent parfois des difficultés, note Reuters, affirmant qu'un des projets de puce passé de Meta a été arrêté, les tests étant décevants. Selon *The Information*, Microsoft aurait aussi repoussé le lancement de son prochain processeur à 2026. *« Nous sommes satisfaits de l'avancement de nos activités dans les puces »*, rétorque Omar Khan, vice-président de sa filiale de cloud. *« Nous sommes engagés dans un projet sur plusieurs générations* [de processeurs] *»*, assure-t-il, affirmant que les puces maison Maia, lancées en 2023, sont utilisées pour certaines requêtes des modèles d'IA de son partenaire OpenAI, ainsi que de Copilot, l'assistant déployé dans sa suite bureautique Office.

Face à la concurrence, Nvidia s'affiche plutôt serein. *« Pourquoi construire une nouvelle puce spécialisée si elle ne va pas être meilleure que celle que vous pouvez acheter* [chez Nvidia] *? »*, a ironisé son patron Jensen Huang, lors d'une conférence à Paris en juin. Ses GPU dernier cri restent techniquement les plus performantes du marché, selon les analystes. Et Nvidia bénéficie de ses investissements importants. D'ailleurs, les géants du cloud restent prudents dans leur communication car ils resteront des clients et partenaires importants de Nvidia.

Le leader des puces pour l'IA peut-il être détrôné ? Les puces maison (ASIC) des acteurs de l'IA et du cloud ont grignoté 20 % de ce marché encore en très forte croissance (soit environ 40 milliards de dollars sur 200 milliards de dollars), pointe M. Sztabowicz. Et, d'ici à fin 2029, leur part pourrait passer à 30 %, selon l'analyste, et même à *« plus de 30 % »*, selon l'institut TrendForce.

En s'implantant dans la conception de puces, les géants de l'IA et du cloud remontent encore la chaîne de valeur pour mieux maîtriser un maillon de plus, après les logiciels d'IA (leurs canaux de distribution auprès du grand public et des entreprises) et les data centers. Et ils contrôlent en partie un composant que les tensions géopolitiques rendent de plus en plus stratégique. Le président Trump a utilisé la restriction de l'exportation de certaines puces consacrées à l'IA dans ses négociations avec la Chine et a menacé fin août de faire de même vers l'Europe.