IA : plongée dans le laboratoire secret d'Amazon, à Austin, où sont créées les puces *du futur*

Les Echos

Il y a dix ans, Amazon a racheté une jeune pousse israélienne, Annapurna Labs, afin de concevoir ses propres semi-conducteurs et de les déployer dans ses centres de données. Ces puces sont désormais un élément clé de sa stratégie dans l'IA et le cloud pour mieux rivaliser avec le californien Nvidia.

Par Hortense Goulard

Dans un immeuble banal de la banlieue d'Austin, au Texas, des ingénieurs travaillent à dessiner et à tester les puces du futur. Il y a dix ans, Amazon a racheté Annapurna Labs, une jeune pousse israélienne, pour 350 millions de dollars. Aujourd'hui, la discrète start-up joue un rôle clé pour l'entreprise fondée par Jeff Bezos.

Même si Amazon continue à être perçue comme une plateforme d'e-commerce, où tout le monde peut commander n'importe quel objet et le voir apparaître, comme par magie, devant sa porte, le géant américain tire en réalité la majeure partie de ses profits de sa branche cloud, AWS (Amazon Web Services).

Pour attirer toujours davantage de clients vers leurs plateformes, les géants du cloud ont commencé à développer leurs propres modèles de puces, spécialisées notamment dans l'entraînement et le fonctionnement des modèles d'IA. Amazon ne fait pas exception à la règle.

Plus de choix

« Nous aimons vraiment l'idée de donner un choix à nos clients », assure Dave Brown, qui dirige les équipes d'AWS consacrées à la puissance de calcul et à l'IA. « Nous voulons nous assurer que lorsqu'un client vient chez AWS, ils ont plusieurs options entre lesquelles ils peuvent choisir pour exécuter leur charge de travail », poursuit le dirigeant.

Amazon est particulièrement attentif au coût et à la performance des puces. Les puces Trainium 2, qui sont conçues spécialement pour effectuer les calculs nécessaires à l'entraînement des modèles d'IA, assurent, selon Dave Brown, un « rapport coût performance qui est de 30 % à 40 % meilleur que des GPU comparables » - les puces H100 de Nvidia.

« Amazon suit la même stratégie que Microsoft et Google, relève Alvin Nguyen, analyste chez Forrester. La seule différence est qu'ils ont fait le choix d'acquérir une expertise externe [en rachetant Annapurna Labs]. Google mise plutôt sur son expertise en interne, tandis que Microsoft a signé des partenariats pour avoir de l'aide sur le design des puces. »

En concevant leurs propres semi-conducteurs, qui sont ensuite produits par TSMC, à Taïwan, les géants du cloud sont capables de les optimiser en fonction de leurs objectifs. « Ils ne veulent pas dépendre d'un seul fournisseur pour leurs cas d'usage les plus importants

», poursuit l'analyste. « Cela leur permet d'être flexibles et proactifs si un nouveau type de modèles d'IA apparaît. »

Tests en laboratoire
Rapidité d'exécution et agilité sont le maître mot des équipes d'Annapurna Labs. Dans le même immeuble, les ingénieurs ont accès à un laboratoire pour tester les dernières générations de semi-conducteurs, ainsi que les circuits imprimés sur lesquels ces puces viennent s'encastrer.

Les équipes travaillent aussi à de nouvelles technologies pour réfrigérer ces puces, pour les connecter entre elles et pour faciliter leur déploiement rapide dans les centres de données.

« Pendant le Covid, j'ai dit au bâtiment que j'avais besoin de laboratoires près de l'endroit où mes ingénieurs travaillaient, raconte en souriant Rami Sinno, qui dirige les équipes d'ingénieurs d'AWS. Ils m'ont dit : 'personne ne va jamais revenir au bureau, vous pouvez avoir autant d'espace que vous voulez.' J'ai doublé la taille de mon laboratoire. »

Intégration verticale
Le directeur est fier de montrer les différentes machines, qui permettent à AWS de tester les capacités des puces en faisant varier la température, la tension ou la fréquence du courant électrique. « Pendant mes trente années d'expérience, je n'ai jamais vu cela : tout est situé dans le même immeuble, du dessin de la puce au rack en passant par le circuit imprimé », souligne-t-il.

« Etre intégré verticalement nous donne un avantage considérable, poursuit le directeur. Nous pouvons décider où investir pour innover, qu'il s'agisse de l'emballage, de la distribution d'électricité, du networking [les câbles qui connectent les puces entre elles]. » L'entreprise passe en revanche commande à des fournisseurs pour d'autres parties, moins cruciales, de ces systèmes compliqués.

Centre de données
Le laboratoire n'est pas le seul endroit où AWS teste ses puces. A quarante-cinq minutes du centre-ville, l'entreprise loue un espace auprès d'un sous-traitant pour tester ses semi-conducteurs dans des conditions réelles, proches de ce qui se fait dans un data center. Cela fait quelques mois seulement que le géant a commencé à déployer ses puces dans cet entrepôt.

A l'intérieur du bâtiment, six racks ont été connectés au réseau. Ils ressemblent à de grandes armoires remplies de fils multicolores, qui alimentent les serveurs en électricité et les relient entre eux. A l'arrière, une « allée chaude » (« hot aisle » en anglais) permet d'évacuer la chaleur. Elle est réfrigérée en permanence par des ventilateurs extrêmement bruyants.

Un peu plus loin, quelques ingénieurs d'AWS s'affairent à connecter deux autres racks. Dans la course à l'IA, la vitesse de déploiement des serveurs dans les centres de données devient un avantage comparatif. Ces derniers sont conçus pour pouvoir être connectés de la façon la plus simple possible au réseau.

Partenariat avec Anthropic
Malgré tous les efforts d'Amazon et d'autres géants du cloud pour se tailler une part du marché des puces IA, Nvidia en reste le leader incontesté. L'entreprise dirigée par Andy Jassy parvient néanmoins à tirer son épingle du jeu en se positionnant sur certaines utilisations de ces puces.

Certains clients choisissent d'utiliser des semi-conducteurs d'AWS parce qu'ils peuvent les adapter facilement à leurs besoins, explique Gadi Hutt, qui dirige les équipes produits d'Annapurna Labs. C'est le cas d'Anthropic, qui a signé un partenariat avec Amazon pour construire Project Rainier, un super-ordinateur utilisant des centaines de milliers de puces Trainium.

« Les grands clients, comme Anthropic et Poolside, peuvent utiliser NKI [une plateforme logicielle] pour optimiser les calculs sur la puce, ajoute le dirigeant. Pour le reste du marché, nous nous concentrons pour l'instant sur les modèles les plus populaires, comme Llama ou DeepSeek. »