



Se préparer à un
futur où l'IA sera
partout



Nul ne connaît aujourd'hui les applications de l'IA dans le futur. Pour s'y préparer, le DSI doit placer son entreprise en capacité d'accueillir ces IA où qu'elles s'exécuteront, que ce soit dans les datacenters, sur le Edge et sur les PC. C'est la position présentée par Intel avec son approche « AI Everywhere » et le lancement de ses AI PC.

Depuis les années 2010, l'IA est sortie de l'ère de glaciation dans laquelle ces techniques étaient plongées depuis plusieurs décennies. Ce réveil s'est matérialisé pour le grand public avec le GPT-3 en 2020, avec une explosion des usages, tant auprès du grand public que des entreprises. Pour autant, le modèle des grands LLM basés dans le Cloud est en train d'évoluer rapidement. L'arrivée des PC accéléré par l'IA laisse augurer un futur où les IA seront beaucoup plus distribuées, une approche formalisée par Intel depuis plusieurs années sous le terme « AI Everywhere ». « Une IA va pouvoir s'exécuter sur n'importe quel type de système, que ce soit le Cloud bien sûr, mais aussi sur les serveurs de l'entreprise, dans l'Edge, les laptops, les objets connectés » explique Tristan Ladrech, Responsable Technique des comptes FSI & Automotive chez Intel Corporation. « Il s'agit d'une véritable nouvelle vague d'IA qui est en train d'arriver. »

L'IA sort enfin du Cloud

De facto, l'approche où toutes les IA étaient exécutées dans des datacenters ou dans le Cloud est en train de faire place à une approche beaucoup plus ouverte et équilibrée. Les industriels peuvent mettre en œuvre des IA au plus près des chaînes de production pour réduire le temps de latence. Les caméras intelligentes peuvent embarquer une IA de reconnaissance de forme dans un « form

« Une IA va pouvoir s'exécuter sur n'importe quel type de système, que ce soit le Cloud bien sûr, mais aussi sur les serveurs de l'entreprise, dans l'Edge, les laptops, les objets connectés. Il s'agit d'une véritable nouvelle vague d'IA qui est en train d'arriver. »

Tristan Ladrech, Responsable Technique des comptes FSI & Automotive chez Intel Corporation.

factor » très restreint. Claude Chauvet, EMEA Technical Sales Specialist, Technologies Poste Client Europe chez Intel souligne que ces usages de l'IA en Edge Computing se développent rapidement grâce à cette approche « AI Everywhere » : « Il est aujourd'hui possible de déployer des IA sur des configurations CPU uniquement, sans GPU, avec des fonctions d'accélération IA des Intel Xeon et Core qui disposent du jeu d'instructions VNNI. Il existe aussi des techniques d'optimisation des réseaux de neurones, avec des accélérations d'un facteur 7 à 20. De fait, il est possible de faire fonctionner des modèles d'IA sur des plateformes compactes dont la consommation énergétique reste faible. »

Demain, chaque collaborateur pourra disposer d'un assistant personnel sur son poste pour l'aider dans son travail quotidien sans les restrictions d'usage liées à la manipulation de données confidentielles ou personnelles. « L'IA générative notamment a pris naissance dans le Cloud tout simplement pour des raisons de puissance de calcul » note Tristan Ladrech, « mais aujourd'hui, avec l'arrivée des PC accélérés par l'IA, la disponibilité de NPU sur tous les terminaux, ce type d'application devient possible sur tous les devices, avec un coût énergétique contenu. »

Des cas d'usage de l'IA potentiellement infinis

Exécuter une IA générative de manière totalement autonome sur le poste client présente d'énormes avantages, notamment en termes de confidentialité et de sécurité des données. L'utilisateur pourra par exemple disposer d'un assistant qui indexera tous ses documents personnels et auquel il pourra demander des analyses, des synthèses sans devoir envoyer les documents concernés dans le Cloud. De même, l'utilisateur pourra interroger son IA pour extraire tous les points importants d'un livre blanc ou d'un dossier technique confidentiel sans risque de fuite de

données. Cette fonction peut être précieuse dans les centres R&D des gros industriels, ou en droit, l'avocat pourra littéralement poser des questions au dossier d'instruction pdf et obtenir ses réponses sans enfreindre le RGPD ou les règles internes liées à la protection des données.

Claude Chauvet ajoute : « Nul ne connaît aujourd'hui tous les cas d'usage de l'IA générative qui apparaîtront dans les prochains mois. Actuellement, les usages se concentrent sur les métiers de la création de contenu, la bureautique et le développement. Les IA vont aussi avoir un gros impact dans l'industrie, notamment les outils de CAO et de simulation numérique. » Pour les industriels, disposer d'IA à toutes les étapes de création des produits va permettre d'accélérer le cycle de conception/développement et réduire le Time to Market de leurs produits.

Un changement de modèle pour les ISV

Tous les ISV travaillent aujourd'hui sur l'intégration de fonctions d'IA dans leurs logiciels. Ceux-ci vont adapter l'architecture de leurs solutions afin de tirer pleinement profit des AI PC qui sont en train d'être déployés chez leurs clients. « Les ISV vont pouvoir embarquer des fonctions d'IA pour un coût beaucoup plus modeste qu'aujourd'hui » estime Claude Chauvet. « Alors qu'ils devaient disposer de fermes de serveurs extrêmement performantes pour exécuter toutes les inférences d'IA de leurs clients, l'AI PC va changer la donne. Avec un modèle distribué sur les postes clients, la réduction du coût sera significative et ils pourront proposer des offres à la fois très performantes pour des coûts d'opération très inférieurs. » Les infrastructures dédiées aux LLM demandent énormément de ressources de calcul, énormément de mémoire. Miser sur des architectures distribuées va apporter une solution à la problématique de montée en charge des serveurs d'IA. Pour l'ISV, cela



représente un coût très important qu'il doit faire passer à ses clients. Une architecture distribuée mettant en œuvre des PC accélérés par l'IA apporte une solution pérenne au problème de montée en charge des modèles d'IA. C'est à la fois une solution en termes de consommation énergétique et d'empreinte CO2, mais aussi une solution à cette course à la puissance sans fin que l'on observe avec l'arrivée des grands LLM.

Cette nouvelle approche va faire sauter quelques freins et faire émerger de nouveaux « business models » pour des modèles d'IA innovants. Une multiplication des cas d'usage des IA est à attendre dans les années à venir, comme c'est déjà le cas dans le domaine de la collaboration et de la visio qui exploite déjà les NPU des processeurs Intel Core Ultra. C'est aussi le cas dans la cybersécurité car avec une puissance de traitement IA bien supérieure au niveau du poste client, les algorithmes comportementaux par exemple

pourront surveiller un nombre de process bien plus important et donc accroître la sécurité des postes.

Déployer des AI PC pour être « Future Proof »

Pour le DSI, faire le choix de standardiser son parc sur des AI PC est la garantie de pouvoir faire fonctionner ces applications au cours des 3 à 5 prochaines années, d'être à l'épreuve du temps vis-à-vis de cas d'usage de l'IA que l'on ne connaît pas encore. « Aujourd'hui, la question n'est plus de savoir si l'entreprise va utiliser de l'IA ou pas » affirme Claude Chauvet. « La vraie question pour le DSI est de choisir les bons services et les applications qui apporteront de la valeur à son entreprise, les IA qui assisteront le mieux les utilisateurs dans leurs tâches. Sur ce plan,

le volet hardware ne doit pas être un frein. »
Que ce soit les Xeon des serveurs ou les Intel Core, les CPU sont aujourd'hui optimisés pour l'IA, les GPU peuvent supporter les workloads d'apprentissage les plus lourdes et les NPU qui permettront de porter des inférences jusque sur le poste de l'utilisateur final.

Pour pouvoir utiliser ces logiciels dopés à l'IA encore en cours de développement, il faudra disposer de toute une palette d'outils et disposer de machines capables de délivrer une performance IA avec une faible consommation énergétique. Aujourd'hui, la différence de coût d'achat entre une machine dotée d'un NPU et d'un CPU classique est négligeable et ne justifie pas le risque pris à ne pas opter pour cette technologie. Un renouvellement de parc représente un investissement sur 3 à 5 ans et le choix de l'AI PC permettra au DSI de garder toute sa liberté de choix pour les futures applications.

« Aujourd'hui, la question n'est plus de savoir si l'entreprise va utiliser de l'IA ou pas »

Claude Chauvet
