ÉTUDE DE MARCHÉ

L’opportunité du PC avec IA

Commandité par intel

RÉSUMÉ

Le vif intérêt suscité par l’IA, en particulier l’IA générative, a créé un engouement sans précédent dans le secteur de la technologie. Toutefois, cet engouement s’est accompagné d’une prise de conscience du fait que les nouvelles capacités offertes par l’IA ne peuvent pas toutes être exploitées dans le Cloud. L’accent a donc été mis sur l’exécution de charges de travail d’IA sur PC et autres appareils clients. Les fournisseurs de semi-conducteurs pour PC, comme Intel, répondent à cet intérêt en proposant de nouvelles puces et de nouveaux logiciels optimisés pour exécuter ce type d’applications. De l’intégration de nouvelles architectures d’accélérateurs d’IA, comme les NPU, dans les SoC de nouvelle génération comme Intel® Core™ Ultra, à une exploitation plus poussée de la puissance des CPU et des GPU, Intel déploie des efforts concertés pour saisir les opportunités offertes par l’exécution d’applications d’IA générative ou d’autres types d’applications d’IA sur les PC. Ces premières initiatives montrent clairement qu’une focalisation simpliste sur les performances en TOPS d’un NPU n’offre pas une évaluation précise de ce que l’expérience de l’utilisation de l’IA sur PC est capable d’accomplir. La prise en charge logicielle dans de nombreux domaines, y compris les outils de développement, les API intégrées dans le système d’exploitation, les Frameworks d’exécution de l’IA, les outils de déploiement et les pilotes au niveau du système, sont tous nécessaires pour tirer parti de tout le potentiel de l’IA sur les appareils clients.

« Ce qui est particulièrement intéressant avec l’IA, c’est qu’elle a permis de faire découvrir à tout un chacun l’étendue des possibilités offertes par les appareils informatiques » – Bob O’Donnell, Analyste en chef.

INTRODUCTION

L'un des effets les plus intéressants de l'engouement suscité par les applications basées sur l'intelligence artificielle (IA) est qu'il a permis de faire évoluer les mentalités quant aux possibilités offertes par les appareils informatiques. Après de nombreuses années de promesses déçues en matière d'IA, le lancement à grande échelle et le succès presque immédiat des modèles de fondation pouvant alimenter les applications d'IA générative telles que ChatGPT d'OpenAI nous ont fait entrer dans une formidable nouvelle ère de l'informatique.

L'IA générative (GenAI) ouvre de nouvelles perspectives en matière d'informatique, de créativité, de productivité, de communication et bien plus encore, inspirant ainsi les utilisateurs du monde entier à tirer pleinement parti de cette technologie. Au-delà de l'IA générative, cette évolution s'étend à de nombreuses applications « traditionnelles » centrées sur l'IA. De l'analyse et du traitement d'images et de vidéos à la productivité bureautique, en passant par la transcription et le résumé de réunions, la modélisation et la texturation 3D, le nettoyage d'images et de vidéos, et bien d'autres choses encore, toutes sortes d'applications alimentées par l'IA connaissent un nouvel essor. En outre, certaines applications « traditionnelles » basées sur l'IA, comme le floutage d'arrière-plan et le débruitage audio, sont également perçues sous un jour nouveau, car elles peuvent tirer parti de certaines ressources informatiques axées sur l'IA qui sont désormais disponibles.

Jusqu'à présent, l'informatique alimentée par l'IA s'est surtout concentrée sur les applications et les services s'exécutant dans le Cloud. Or, il est désormais possible d'exécuter ce type d'applications directement sur PC et autres appareils clients. D'ailleurs, outre le fait que cette possibilité se développe rapidement, dans certaines situations, les performances et les résultats obtenus peuvent être meilleurs lorsque l'application est exécutée localement. En outre, le fait de pouvoir exploiter les données sur votre propre appareil et de ne pas les envoyer dans un environnement de Cloud public présente des avantages considérables en termes de confidentialité et de sécurité.

Certaines de ces nouvelles possibilités sont le fruit des progrès considérables accomplis par les solutions d'IA et d'IA générative embarquées au cours des derniers mois. L'évolution rapide et la rationalisation des modèles de base open-source, ainsi que les avancées technologiques telles que la quantification des modèles, vont permettre de réaliser au cours des prochains mois des choses que de nombreux observateurs du secteur ne s'attendaient pas à retrouver sur les appareils clients avant plusieurs années. D'ailleurs, au cours des derniers mois, le rythme de l'innovation sur les appareils a même été plus rapide que celui des avancées de l'IA générative, ce qui n'est pas peu dire !

Outre des avancées technologiques impressionnantes, ces progrès spectaculaires de l'IA sur les appareils s'expliquent en grande partie par des questions très pratiques. Il est en effet largement admis que, compte tenu de l'incroyable rapidité d'adoption des outils d'IA générative et de la multitude de nouvelles solutions proposées en ligne, l'infrastructure existante des centres de données dans le Cloud public ne peut tout simplement pas répondre à l'ensemble de la demande prévue. De plus, les besoins en énergie qu'exigeraient ces ressources basées sur le Cloud suscitent d'importantes inquiétudes. Enfin, les considérations de coût, de sécurité et d'efficacité laissent à penser que l'exécution de toutes les charges de travail d'IA, ou même de la plupart d'entre elles, dans le Cloud n'est tout simplement pas envisageable à long terme. Par conséquent, les solutions d'IA embarquées sur les appareils deviennent essentielles pour garantir que la dynamique autour des applications alimentées par l'IA puisse se poursuivre. Davantage de ces charges de travail d'IA doivent tout simplement être transférées sur les PC.

L'IMPORTANCE D'UN SOC EQUILIBRE

Dans ce contexte, une attention particulière est accordée aux types d'applications et de charges de travail d'IA qui peuvent s'exécuter sur des appareils clients tels que les PC. Cette question est évidemment directement liée au type de ressources informatiques que les PC modernes peuvent offrir. Comme nous le verrons un peu plus loin dans ce document, il convient également de s'intéresser de plus près aux logiciels nécessaires pour tirer parti de ce matériel informatique.

Cette année a été marquée par le lancement de plusieurs nouvelles architectures de systèmes sur puce (System on Chip, ou SoC) pour PC, qui peuvent être exploitées pour exécuter des charges de travail d'IA de manière plus puissante et plus efficace que les itérations précédentes. Ainsi, des puces comme le nouveau SoC Intel® Core™ Ultra (anciennement connu sous le nom de « Meteor Lake ») sont désormais dotées de processeurs plus flexibles, de processeurs graphiques plus puissants et d'un nouveau type de composant appelé NPU (unité de traitement neuronal) qui est spécifiquement optimisé pour de nombreux types de charges de travail d'IA.

L'intégration du NPU, en particulier, suscite beaucoup d'intérêt pour ces nouvelles architectures SoC modernes et pour les possibilités qu'elles offrent en matière d'intelligence artificielle sur l'appareil. Les NPU visent à accélérer les performances de multiplication de matrices et d'autres équations mathématiques qui sont souvent utilisées par des applications alimentées par l'IA ou par des fonctions au sein d'applications plus vastes. Elles permettent surtout d'améliorer les performances des algorithmes et autres composants logiciels qui tournent en permanence en arrière-plan, comme le font souvent les applications d'assistants numériques et autres « agents intelligents ».

Les NPU, aussi puissants soient-ils pour certains types d'applications d'IA, ne constituent pas une solution miracle pour toutes les applications d'IA. De fait, la plupart des travaux d'inférence IA effectués sur les PC sont toujours réalisés avec des processeurs et d'autres types de calculs liés aux charges de travail IA peuvent être effectués plus efficacement par des processeurs graphiques. Le point le plus important à retenir est que pratiquement tous les processus alimentés par l'IA qui s'exécutent sur un PC peuvent être traités par n'importe lequel des divers composants architecturaux d'un SoC (CPU, GPU ou NPU), les niveaux d'efficacité étant simplement différents d'un composant à l'autre. De plus, sur une puce comme Intel® Core™ Ultra qui comporte à la fois des P-cores plus performants mais gourmands en énergie et des E-cores aux performances moindres mais plus économes en énergie, il arrive qu'un type de processeur soit meilleur que l'autre pour une charge de travail d'IA donnée, et vice-versa. D'une manière générale, les processeurs sont utilisés pour l'IA légère à inférence unique et à faible latence, les processeurs graphiques pour les charges de travail exigeantes en matière d'IA, et les NPU pour l'IA continue et le délestage de l'IA.

Il s'agit simplement de choisir le bon outil pour le bon travail. Comme beaucoup d'entre nous l'ont sans doute appris sur le terrain, un marteau permet de faire beaucoup plus de choses que ce pour quoi il a été conçu à l'origine, mais certains outils rendent certaines tâches beaucoup plus faciles (et plus rapides !).

À cet égard, outre les nouvelles capacités, il est également important de discuter des indicateurs de performance et d'efficacité lorsqu'il s'agit d'exécuter des applications d'IA sur des PC. L'indicateur le plus couramment évoqué par de nombreuses entreprises est le TOPS, ou téra-opérations par seconde, un mécanisme de mesure conçu à l'origine pour mesurer les calculs mathématiques. On s'est notamment beaucoup intéressé au TOPS du NPU d'un système.

Il s'avère que ce n'est pas le meilleur indicateur lorsqu'il s'agit de mesurer l'expérience en situation réelle, et ce pour plusieurs raisons. Tout d'abord, pour beaucoup, le TOPS ne reflète pas vraiment les performances réelles : il s'agit davantage d'un indicateur synthétique simpliste que d'une mesure perceptible en situation réelle. En effet, le TOPS mesure le nombre de calculs effectués, et non le type de calculs, qui a souvent un impact beaucoup plus important sur les performances réelles. Un autre indicateur, le TOPS/watt, évalue l'efficacité globale sur la base de la consommation d'énergie lors de l'exécution de certains calculs. Bien que cet indicateur soit généralement perçu comme plus pertinent, il ne s'agit pas d'un point de comparaison idéal.

L'autre gros défaut est que ces indicateurs ne tiennent pas compte du fait que, comme nous l'avons mentionné plus haut, les charges de travail IA peuvent s'exécuter (et s'exécutent souvent) sur plusieurs composants différents d'un système PC complet. C'est la raison pour laquelle on parle davantage du TOPS général du système, qui combine les TOPS potentiels du

processeur, du processeur graphique et du NPU afin d'obtenir un chiffre qui reflète davantage les systèmes qui exécutent différents types d'applications d'IA.

Cependant, même le TOPS du système n'est pas optimal, car il ne tient pas toujours compte de l'expérience acquise lors de l'utilisation de différentes applications. En outre, un autre problème majeur lié à la mesure des performances de l'IA sur un PC est que les indicateurs traditionnels basés sur la vitesse pure n'ont pas vraiment de sens lorsqu'il s'agit d'IA. Si certains peuvent s'intéresser à la rapidité avec laquelle un chatbot doté d'un grand modèle de langage (Large Language Model, ou LLM) tournant localement peut répondre à votre demande, par exemple, ce n'est pas le cas du plus grand nombre. Des éléments tels que la qualité de la réponse et l'impact sur l'autonomie de la batterie auront probablement une incidence bien plus importante sur l'opinion des utilisateurs quant aux performances d'un PC avec IA par rapport à un autre.

Et pour compliquer encore un peu plus l'évaluation comparative des PC avec IA, il s'avère que d'autres facteurs peuvent influer beaucoup plus sur les performances que le TOPS d'un composant donné (ou même du système dans son ensemble). Pour de nombreux modèles LLM, en raison de la taille des ensembles de données qu'ils utilisent, la quantité de mémoire système et la vitesse d'accès à cette mémoire ont une incidence beaucoup plus importante sur les performances réelles que le TOPS à tous les niveaux. En d'autres termes, un accès plus rapide à une plus grande quantité de mémoire sur des systèmes dont les performances TOPS sont inférieures peut offrir de meilleures performances en situation réelle que d'autres systèmes dont les spécifications TOPS sont supérieures.

OUTILS LOGICIELS D'IA

S'il est important de disposer de capacités matérielles avancées, il n'en reste pas moins que, comme pour tout ce qui touche à l'informatique, ces capacités ne servent à rien si l'on ne dispose pas des outils logiciels adéquats. Les modèles, algorithmes et Frameworks de développement de l'IA/ML/DL s'avèrent particulièrement importants car les logiciels basés sur l'IA sont en pleine évolution.

Ainsi, comme nous l'avons évoqué précédemment, d'énormes progrès ont été réalisés pour faire évoluer les modèles de base extrêmement volumineux basés sur le Cloud qui ont contribué à alimenter le monde des applications et des services d'IA générative vers un format qui s'adapte et s'exécute nativement sur les PC.

Des versions réduites de grands modèles comme le Llama 2 de Meta, le nouveau Gemini de Google et bien d'autres offrent la possibilité d'exécuter des modèles de base avec moins de 10 milliards de paramètres directement sur les PC sans avoir besoin d'une connexion au Cloud.

La multitude de modèles open-source et la croissance des places de marché telles que Hugging Face offrent également aux développeurs la possibilité d’élaborer des modèles spécifiquement conçus pour s’exécuter sur des appareils clients. En outre, de nombreux progrès ont été réalisés récemment pour quantifier des modèles de grande taille afin de les adapter aux limites des ressources d’un PC. Ensemble, ces avancées et les nombreuses autres à venir font rapidement passer l’IA embarquée du stade de la science-fiction à court terme à la réalité scientifique en temps réel.

Dans le cas des charges de travail IA exécutées sur PC, divers composants logiciels au niveau du système et des applications jouent également un rôle très important dans la qualité de l’expérience de l’IA embarquée sur un PC par rapport à un autre. Sur les PC sous Windows, par exemple, certains éléments du système d’exploitation jouent un rôle essentiel dans la répartition des charges de travail et des fonctions basées sur l’IA entre les différents composants du matériel d’un système. DirectML, en particulier, joue le rôle de « régulateur de trafic » pour les applications basées sur l’IA dans Windows, en aiguillant divers éléments logiciels ou sous-programmes d’une application donnée vers le meilleur élément matériel sur le SoC d’un PC. La version la plus récente de DirectML intègre un certain nombre d’améliorations logicielles qu’Intel a spécifiquement créées et fournies à Microsoft afin d’améliorer l’écosystème logiciel général pour les applications basées sur l’intelligence artificielle sur les PC (y compris les systèmes utilisant les SoC d’autres fournisseurs). Des éléments tels que DirectML jouent un rôle très important dans l’amélioration des performances en situation réelle, lorsque plusieurs applications ou agents alimentés par l’IA s’exécutent simultanément. Dans ce cas, il est essentiel d’équilibrer la combinaison des différents composants logiciels fonctionnant sur différents éléments matériels du SoC afin d’obtenir des résultats différents en termes de puissance et de performances.

Outre ces améliorations au niveau du système, pour tirer les meilleures performances possibles d’une application donnée, il convient de travailler directement avec les développeurs de logiciels afin de s’assurer que leur code est optimisé pour une architecture donnée. C’est là que l’envergure d’Intel et son large éventail de développeurs de logiciels internes lui confèrent souvent un avantage en raison de sa capacité à atteindre et à collaborer avec un grand nombre d’éditeurs de logiciels indépendants (ISV) travaillant sur des applications PC alimentées par l’IA. Dans la même veine, Intel a lancé une nouvelle initiative sur les logiciels d’IA, qui consiste à collaborer avec les 100 principaux développeurs d’applications orientées IA afin de s’assurer que leurs applications fonctionnent le plus efficacement possible sur le silicium d’Intel.

Enfin, le dernier volet de la question des logiciels, souvent négligé, concerne les outils de développement. Les développeurs de logiciels utilisent souvent les outils des fournisseurs de processeurs pour créer leurs applications. Un environnement de développement comme OpenVINO™ d’Intel peut grandement contribuer à accélérer et à enrichir leur travail.

OpenVINO™ est assorti d'un ensemble de plus de 200 modèles d'intelligence artificielle pré-entraînés qui ont été construits et testés pour s'exécuter sur des PC (et de manière plus efficace sur les SoC d'Intel). En outre, OpenVINO™ comprend une API de conversion de modèles qui permet aux développeurs d'intégrer de nouveaux modèles publics ou open-source dans OpenVINO™, ce qui leur donne plus de flexibilité et de possibilités lorsqu'il s'agit de construire leurs applications basées sur l'IA. OpenVINO™ prend également en charge les modèles formés avec Pytorch et TensorFlow et sert de backend intégré pour Hugging Face Optimum et torch.compile de Pytorch, offrant ainsi aux développeurs d'applications un large éventail de possibilités.

APPLICATIONS D'IA SUR PC

Nous commençons déjà à voir un certain nombre d'applications pour PC et de fonctions au niveau du système qui exploitent les fonctionnalités de l'IA. Les fonctions Windows Studio Effects de Microsoft, qui ont été spécifiquement optimisées pour s'exécuter sur les NPU des PC qui en sont équipés, permettent d'améliorer le flou de l'arrière-plan vidéo et la réduction du bruit audio dans les fonctions de messagerie en temps réel. Mieux encore, elles s'exécutent de manière nettement plus efficace que si elles s'exécutaient sur le processeur ou le processeur graphique d'un PC.

Les possibilités offertes par un outil comme Rewind.ai, dont Intel a fait la démonstration lors de son récent événement sur l'innovation, sont particulièrement intéressantes. Comme son nom l'indique, Rewind.ai consigne tout ce que vous faites et dites avec votre PC, des e-mails aux documents en passant par les chats, les réunions en ligne et bien plus encore, et vous propose des résumés alimentés par l'IA générative ainsi qu'un accès à toutes ces informations. Cela préfigure le type d'assistant véritablement numérique dont beaucoup d'entre nous ont rêvé depuis les premiers outils, beaucoup moins puissants (et utiles), comme Cortana, Siri, Alexa et d'autres.

Dans un tout autre registre, les versions les plus récentes de Lightroom d'Adobe et de Magix de Vegas intègrent la technologie d'amélioration des images et des vidéos par l'IA générative et peuvent utiliser le NPU local d'un PC pour accélérer les opérations. Nous avons également commencé à voir apparaître d'autres outils de génération d'images alimentés par l'IA générative qui n'ont pas besoin d'une connexion basée sur le Cloud pour fonctionner. Comme pour toute application s'exécutant localement, cela améliore grandement la confidentialité et la sécurité lors de l'utilisation de ces applications, car aucune de vos informations ne peut être capturée dans le Cloud.

Il est également important de réfléchir aux conséquences que les récentes réductions de la taille des modèles LLM pourraient avoir sur des applications PC encore plus courantes. Alors que les dernières suites de productivité M365 de Microsoft et Workspace de Google exploitent toutes deux le Cloud pour la plupart de leurs fonctionnalités d'IA générative, la perspective d'exécuter certaines de ces fonctions directement sur le PC avec ces petits modèles LLM se rapproche à grands pas. De plus, la possibilité de personnaliser ces LLM avec vos propres données (ou celles de votre entreprise) est particulièrement intéressante. Cette possibilité d'effectuer des personnalisations plus poussées sur la base de données stockées localement ou sur l'intranet de l'entreprise peut aboutir à la création d'outils encore plus performants et mieux ciblés que tout ce qui est accessible dans le Cloud. En outre, ces opérations peuvent être effectuées plus rapidement si toutes les données utilisées sont stockées localement sur l'appareil. C'est d'ailleurs probablement l'une des raisons les plus déterminantes et les plus puissantes pour lesquelles les applications d'IA embarquées sur l'appareil présentent autant d'intérêt.

Une autre piste intéressante que les entreprises commencent à explorer consiste à tirer parti du concept d'IA hybride, où certains aspects du travail s'effectuent dans le Cloud et d'autres sur le PC. Imaginons par exemple un scénario selon lequel un programme de retouche d'images crée une version à plus faible résolution adaptée à l'écran sur le PC, mais crée ensuite séparément une version à plus haute résolution via un modèle basé sur le Cloud. La version à faible résolution peut être modifiée rapidement sur le PC, mais c'est la version basée sur le Cloud qui sera finalement sauvegardée. Dans un environnement professionnel tel que le secteur très réglementé de la santé, il y a également eu quelques exemples précoces d'entreprises générant des e-mails personnalisés sur des procédures médicales à l'aide de plusieurs modèles. Dans ces cas, les données à caractère personnel sont traitées sur un modèle local sur le PC, tandis que les éléments plus génériques de l'e-mail sont générés à l'aide d'un grand modèle LLM basé dans le Cloud. L'assemblage du message final se fait ensuite en fusionnant ces deux éléments dans l'e-mail généré. Ces exemples, et bien d'autres que nous verrons probablement en 2024, mettent en évidence la fluidité de l'utilisation de l'IA générative à l'avenir. Ils montrent également que le PC jouera un rôle beaucoup plus important dans l'IA générative qu'il n'y paraît à première vue.

Bien entendu, certaines applications sur PC nécessiteront de puissants NPU pour fonctionner efficacement (ou même pour fonctionner tout court), mais la grande majorité des applications conçues pour tourner sur PC exploiteront le NPU en tant qu'accélérateur, s'il est disponible. Le concept est similaire au rôle joué par les processeurs graphiques. Sur les systèmes dotés de processeurs graphiques discrets plus puissants, certaines fonctions peuvent s'exécuter plus rapidement ou certains jeux peuvent tourner dans des modes de résolution plus élevés ou à des fréquences d'images plus rapides que sur les systèmes dotés de solutions graphiques

intégrées, mais dans la plupart des cas, ils s’exécutent quand même. Au fil du temps, à mesure que des NPU plus puissants seront déployés sur un plus grand nombre de PC, les développeurs de logiciels tireront certainement davantage parti de ces capacités. Mais comme pour la plupart des avancées technologiques, il faut du temps pour que ces évolutions portent leurs fruits.

CONCLUSIONS

Il est indéniable que l’IA générative et l’IA de manière générale ont ouvert de nouvelles perspectives quant aux possibilités offertes par nos appareils informatiques. Alors que la plupart des gens se résignaient à devoir tirer parti d’une connexion basée sur le Cloud pour accéder à la puissance de ces outils, il apparaît clairement que l’IA embarquée n’est pas seulement possible, elle est nécessaire. Et, dans un avenir relativement proche, elle devrait offrir une expérience encore meilleure que celle proposée actuellement dans le Cloud.

Toutes ces raisons, et bien d’autres encore, justifient l’enthousiasme général que suscite l’évolution de l’informatique. Comme beaucoup l’ont dit, il s’agit d’une occasion unique de commencer à travailler d’une manière nouvelle et passionnante.

Et malgré les réserves initiales, il est désormais évident que les PC vont jouer un rôle extrêmement important dans ces efforts à venir. Qu’il s’agisse d’avancées passionnantes dans les architectures de silicium ou d’évolutions importantes dans les applications et outils logiciels basés sur PC, le PC est sans doute au seuil d’une renaissance novatrice et passionnante. Certes, il subsiste un certain nombre de questions concernant les moyens de mesurer au mieux les améliorations de performances que nous sommes sur le point d’observer et il est permis de penser qu’il est peut-être temps de changer complètement notre façon d’envisager les benchmarks et autres indicateurs de mesure.

Quelle que soit la réponse qui sera apportée à ces questions, l’industrie du PC vit une période passionnante, qu’il convient d’apprécier à sa juste valeur.