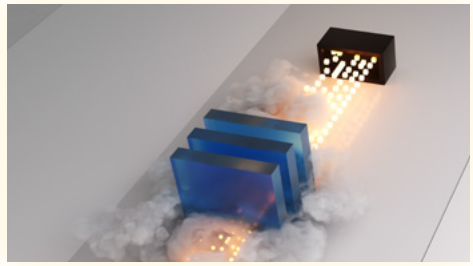
Innovez **plus rapidement** grâce à l'**IA intégrée**

Du concept à la production : transposer l'intelligence artificielle (IA) à grande échelle constitue un défi. Rendre l'IA opérationnelle sur l'ensemble du pipeline de bout en bout, que ce soit en local, dans le Cloud ou en utilisant une approche hybride, impliquait bien souvent des dépenses supplémentaires auxquelles s'ajoutait la difficulté de recruter les compétences adéquates.

Pourquoi l'IA est-elle si **complexe ?**

Le gouffre entre le concept et la production

Matériel spécialisé, compétences avancées et outils personnalisés



Complexité, coût et risque

Pour les chefs d'entreprise qui cherchent à étendre l'IA à l'ensemble de leurs activités, il est essentiel d'en réduire la complexité. Les entreprises veulent pouvoir s'adapter, réduire les coûts et fournir de nouveaux services, et il est donc plus que jamais essentiel que la technologie leur apporte une valeur opérationnelle. Plutôt que de personnaliser les systèmes en fonction des nouvelles applications, ce qui ajoute une couche de complexité supplémentaire, les entreprises peuvent obtenir les performances dont elles ont besoin pour répondre à une grande variété de déploiements actuels et futurs, grâce à une plateforme évolutive.

CADRES SUPÉRIEURS ESTIMANT QU'ILS ONT BESOIN DE L'IA POUR RÉUSSIR
84 %[1]

Démocratiser l'IA pour que tout le monde en profite

Nous sommes à un moment charnière avec les technologies d'IA. La nouvelle ère de l'IA doit permettre à l'ensemble de l'écosystème de bénéficier d'un accès, de visibilité, de transparence et de confiance. Tout cela ne sera possible qu'avec une approche alignée sur les normes de l'industrie, focalisée sur la rentabilité, et mettant fin aux cloisonnements.

APPLICATIONS D'ENTREPRISE QUI UTILISERONT UNE IA INTÉGRÉE D'ICI 2025
90 %[2]

L'IA commence avec Intel

Inférence jusqu'à
1,8x
plus rapide qu'un processeur AMD EPYC 96 cœurs[3]

Intel a conçu des systèmes prêts à être déployés et des outils de développement optimisés et prêts à fonctionner sur les plateformes de serveurs d'inférence d'IA les plus répandues.

Pour ceux désirant repousser les limites, le portfolio d'accélérateurs spécialisés d'Intel est incomparable. Les processeurs de formation et d'inférence d'apprentissage profond Habana, les VPU basse consommation, les FPGA programmables et les processeurs graphiques discrets d'Intel étoffent le socle basé sur le processeur avec des capacités permettant de répondre à une grande variété de charges de travail.

- Les processeurs Intel® Xeon® Scalable de 4e génération intègrent un très grand nombre d'accélérateurs pour les charges de travail clés, y compris l'IA, ainsi que des outils de science des données, et un écosystème de solutions intelligentes. Les solutions Intel sont optimisées pour les charges de travail d'entreprises, de Cloud, de calcul haute performance, de réseau, de sécurité ou encore d'internet des objets, avec des cœurs puissants et une large gamme de fréquences, de fonctionnalités, et de niveaux de puissance.
- Intel® AMX (Intel® Advanced Matrix Extensions) accélère les capacités d'IA sur les processeurs Intel® Xeon® Scalable de 4e génération, accélérant l'entraînement et l'inférence de l'apprentissage profond sans matériel supplémentaire. Cet accélérateur est idéal pour le traitement du langage naturel, les systèmes de recommandation et la reconnaissance d'images.
- Intel® SGX (Intel® Software Guard Extensions) **est actuellement une des technologies d'informatique confidentielle les plus étudiées, actualisées et déployées dans les centres de données, avec l'une des plus petites limites de confiance parmi toutes les technologies d'informatique confidentielle du centre de données.**
- Intel® AVX-512 (Intel® Advanced Vector Extensions 512) peut accélérer le prétraitement des données non structurées provenant de sources multiples pour les modèles d'entraînement, tout en accélérant le mouvement des données pour réduire le temps de traitement des ensembles de données. Le kit d'extension Intel pour Scikit-learn, associée à Intel® AVX-512, accélère également les algorithmes d'apprentissage automatique pour l'entraînement et l'inférence.
- Intel® DL Boost (Intel® Deep Learning Boost) est intégré pour exécuter des charges de travail d'IA complexes sur le même matériel que vos charges de travail existantes. Les canaux Intel® UPI (Intel® Ultra Path Interconnect) augmentent l'évolutivité de la plateforme et améliorent la bande passante interprocesseurs pour les charges de travail gourmandes en E/S.
- L'accélérateur Habana® Gaudi®2 offre un entraînement et une inférence d'apprentissage profond à haute performance et haute efficacité. Il est particulièrement bien adapté à la complexité de l'IA générative et aux modèles de langage de grande taille.

Intel® Security Solution for Fortanix Confidential AI Platform

Fortanix®

- Les modèles et données d'IA peuvent être partagés sans exposer la propriété intellectuelle ni les données sensibles.
- Solution de sécurité haute performance clé en main, de niveau entreprise, ne nécessitant pas de modification des applications.
- Répond aux préoccupations liées au délai de mise sur le marché en fournissant une solution validée avec un guide d'installation, des outils conteneurisés et des charges de travail types.

La solution ouvre la voie au déploiement d'applications d'IA non modifiées dans des enclaves sécurisées au sein de n'importe quel environnement Cloud. Elle s'appuie sur trois piliers intégrés : Fortanix Confidential AI, Fortanix Confidential Computing Manager et les processeurs Intel® Xeon® Scalable de 4e génération avec Intel® SGX (Intel® Software Guard Extensions).[5] En tant que service clé en main, elle permet un provisionnement immédiat prêt à l'emploi sans nécessiter de compétences approfondies en matière d'IA. Cette solution préserve également l'intégrité des modèles en sécurisant les données et les modèles dans des enclaves sécurisées. Cela simplifie la protection à toutes les étapes du cycle de vie de la sécurité des données, de sorte que les entreprises peuvent facilement être opérationnelles avec l'IA.

Performances d'inférence en apprentissage profond jusqu'à 7,78x[4]

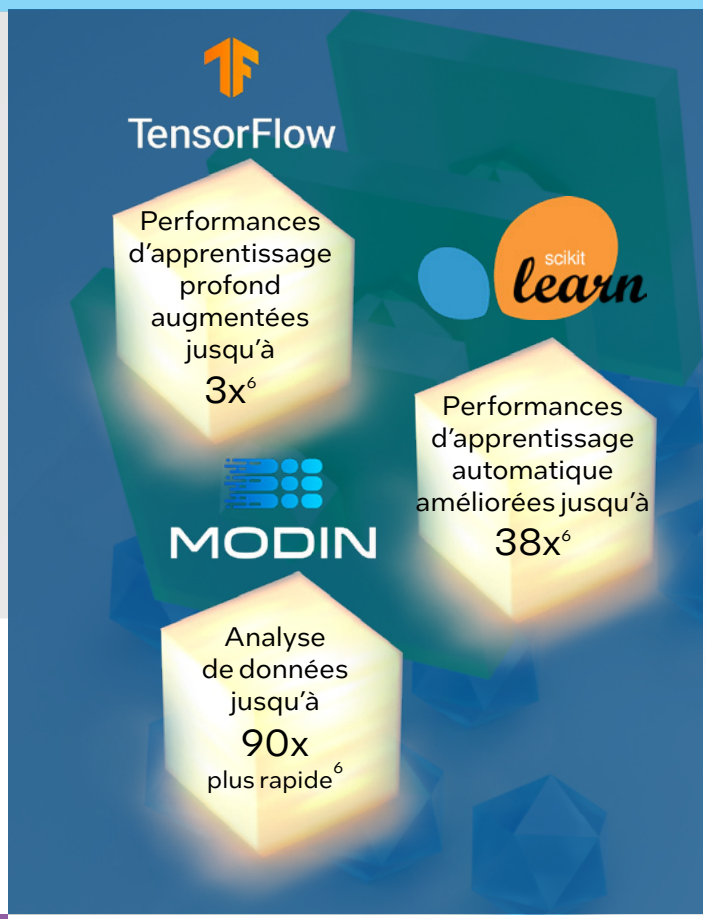
En savoir plus

Notre engagement à intégrer l'IA dans nos plateformes et à la maintenir **ouverte**

Nous nous engageons auprès de nos clients à intégrer l'IA dans nos plateformes et à la maintenir ouverte en optimisant les logiciels en amont dans les environnements de développement d'IA et apprentissage automatique afin de promouvoir la programmabilité, la portabilité et l'adoption par l'écosystème. L'IA en tant que charge de travail doit être accessible et intégrée à différentes applications. Dans cette perspective, nous pensons offrir le choix et la compatibilité entre les architectures, les fournisseurs et les plateformes Cloud, à l'appui d'un écosystème ouvert d'informatique accélérée.

Nos plateformes logicielles et nos architectures hétérogènes composées de processeurs, de processeurs graphiques et d'accélérateurs d'apprentissage profond facilitent le déploiement rapide de l'IA à l'échelle, du Cloud au réseau, jusqu'à la périphérie et au client. Notre pipeline d'IA de bout en bout permet aux développeurs de n'écrire qu'une seule fois et de déployer leur code n'importe où, et nous fournissons un modèle de programmation unifié qui permet de passer facilement d'un déploiement basé sur Intel® Xeon® à nos processeurs graphiques, ainsi qu'à nos accélérateurs dédiés. Au fondement de nos plateformes et de nos solutions se trouve la confiance. Les clients peuvent sécuriser diverses charges de travail d'IA dans le centre de données et l'inférence à la périphérie avec l'informatique confidentielle.

Les solutions qui fonctionnent avec la technologie Intel permettent de faire évoluer l'IA. Les optimisations Intel® AI pour Spark, TensorFlow, PyTorch, scikit-learn, NumPy et XGBoost apportent des gains de performance substantiels[6], tandis que la Distribution Intel® du kit d'outils OpenVINO™ facilite le déploiement de l'inférence de l'apprentissage profond pour des centaines de modèles pré-entraînés. Et les kits d'outils Intel® oneAPI permettent une réutilisation maximisée du code à travers les piles et les architectures, avec la possibilité pour les accélérateurs tels qu'Intel® AMX de fonctionner avec peu ou pas de modifications du code.[7] Pour en savoir plus, rendez-vous sur www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/



Logiciel libre : principales optimisations d'Intel® pour l'IA

PyTorch
De la recherche à la production, développez et déployez des modèles en utilisant les optimisations Intel® dans l'environnement de développement PyTorch par défaut et Intel® Extension for PyTorch.

TensorFlow
Accélérez l'entraînement et l'inférence avec des optimisations par défaut en amont et une extension pour les améliorations de performance les plus récentes.

Scikit-learn
Accélérez vos flux de travail classiques d'apprentissage automatique scikit-learn en changeant deux lignes de code.

Modin
Mettez à l'échelle vos flux de travail Pandas en modifiant une seule ligne de code.

Kit d'outils Intel® AI Analytics
Accélérez la science des données de bout en bout et les pipelines d'apprentissage automatique à l'aide d'outils et d'environnements de développement basés sur Python.

Distribution Intel® du kit d'outils OpenVINO™
Déployez des applications d'inférence hautes performances des appareils vers le Cloud.

BigDL
Pour la formation ou l'inférence distribuée, faites évoluer vos modèles d'IA de manière transparente vers des groupes de mégadonnées composés de milliers de nœuds.

Intel® Neural Compressor
Accélérez l'inférence de l'IA en limitant l'impact sur la précision grâce à la compression automatisée des modèles.

Meituan accélère les services d'inférence de l'IA pour la vision, et optimise les coûts

Meituan souhaitait améliorer le débit de son inférence d'IA pour la vision sans compromettre la précision, afin de prendre en charge des opérations plus intelligentes. Bien que les processeurs graphiques discrets puissent répondre aux exigences de performance, leurs prix sont parfois relativement élevés. Pour les services d'inférence de modèles de longue traîne à faible trafic, les processeurs sont parfois plus rentables. Afin d'accélérer l'inférence de l'IA, Meituan utilise des capacités matérielles avancées telles que les processeurs Intel® Xeon® Scalable de 4e génération et l'accélérateur intégré Intel® AMX (Intel® Advanced Matrix Extensions).

美团

Des performances d'inférence
jusqu'à 4,13x supérieures,
en convertissant les modèles de FP32 à BF16[8]

Amélioration par 3x
avec l'efficacité globale des ressources en ligne et économies de 70 % sur les coûts de service[8]

Hugging Face

INFÉRENCE EN TEMPS RÉEL ACCÉLÉRÉE 5,7x[9]

AIBLE

RÉDUCTION DU TEMPS D'ÉVALUATION DES DONNÉES de plusieurs semaines à 10 MINUTES PAR ENSEMBLE DE DONNÉES[10]

Numenta

DÉBIT D'INFÉRENCE amélioré par 62x[11]

Accélérateurs intégrés aux processeurs Intel® Xeon® Scalable

Afin d'activer de nouvelles fonctions d'accélérateur intégrées dans un environnement hyper dimensionné, Intel aide l'écosystème avec des API Cloud, des bibliothèques, ainsi que des logiciels au niveau du système d'exploitation les plus courants. Il en résulte une utilisation plus efficace des processeurs, et un meilleur retour sur investissement des services.

Avec Intel, les entreprises peuvent accélérer leur déploiement avec l'un des plus grands écosystèmes de partenaires. Les fournisseurs de matériel et de logiciels, ainsi que les intégrateurs de solutions construisent leurs produits en utilisant des processeurs Intel® Xeon® Scalable, offrant un grand choix et une importante interopérabilité avec l'assurance de milliers d'implémentations concrètes.

L'IA est une technologie incroyablement puissante, au potentiel inestimable. Mais elle est encore relativement immature. Nous voulons veiller à ce que la technologie de l'IA progresse de manière responsable. L'industrie, le monde académique, ainsi que les dirigeants doivent travailler ensemble pour façonner notre avenir technologique, et créer de nouvelles perspectives.

Invoquez l'avenir de l'IA

Complément d'infos

Processeurs Intel® Xeon® Scalable de 4e génération

IA et apprentissage profond Intel®

Informations sur l'IA Intel®

[1]Accenture, 19 novembre 2019. « AI: Built to Scale. » https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments.

[2]Grand View Research, Rapport d'analyse sur la taille, les parts et les tendances du marché de l'intelligence artificielle par solution, par technologie (apprentissage profond, apprentissage automatique, traitement du langage naturel, vision artificielle), par utilisation, par région, et prévisions sectorielles, 2022 - 2030. (« Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, Natural Language Processing, Machine Vision), By End Use, By Region, And Segment Forecasts, 2022 - 2030. ») https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

[3]Charge de travail du système de recommandation DLRM, type de données BF16. Voir https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalableprocessors/.

[4]Voir les configurations pour plus de détails.

DÉTAILS DE CONFIGURATION

TEST-1 : Test effectué par Intel le 21 novembre 2022. 1 nœud, 2x Intel® Xeon® Platinum 8380 processeur @ 2,30 GHz, 40 cœurs, HT désactivé, Turbo activé, Mémoire totale 512 Go (16x32 Go DDR4 3200 MT/s [exécution @3200 MT/s]), BIOS version SE5C6200.86B.0022 .D64.2105220049, ucode version 0xd000375, OS Version Ubuntu 22.04.1 LTS, noyau version 6.0.6-060006-generic, workload/benchmark Inférence d'apprentissage profond dans des enclaves sécurisées avec Fortanix, framework version TensorFlow 2.11, nom & version du modèle ResNet50v1.5/Bert-Large.

TEST-2 : Test effectué par Intel le 21 novembre 2022. 1 nœud, 2x Intel® Xeon® Platinum 8480+ processeur @ 2,0 GHz, 56 cœurs, HT désactivé, Turbo activé, Mémoire totale 512 Go (16x32 Go DDR5 4800 MT/s [exécution @4800 MT/s]), BIOS version 3A05, ucode version 0x2b000070, OS Version Ubuntu 22.04.1 LTS, noyau version 6.0.6-060006-generic, workload/benchmark Inférence d'apprentissage profond dans des enclaves sécurisées avec Fortanix, framework version TensorFlow 2.11, nom & version du modèle ResNet50v1.5/Bert-Large.

[5]Intel® SGX n'est pas vulnérable à la plupart des menaces au niveau du système d'exploitation, et la base de données contient aujourd'hui plus de 140 000 menaces : https://cve.mitre.org.

[6]Voir intel.com/content/www/us/en/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html pour les charges de travail et les configurations. Vos résultats peuvent varier.

[7]Voir la documentation technique pour les versions logicielles qui prennent en charge les accélérateurs.

[8]Pour des informations plus complètes sur les performances et les résultats des benchmarks, rendez-vous sur https://www.intel.com/content/www/us/en/customer-spotlight/stories/meituan-vision-ai-customer-story.html.

[9]Voir la déclaration [A2] à la page https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/.

[10]Consultez https://enaible.aible.com/hubfs/Brochure_Aible_Intel_Overstock.pdf pour en savoir plus.

[11]Voir [P6] à la page intel.com/processorclaims : processeurs Intel® Xeon® Scalable de 4e génération. Vos résultats peuvent varier.

Avis et avertissements

Les performances varient selon l'usage, la configuration et d'autres facteurs. Rendez-vous sur www.Intel.com/PerformanceIndex.

Les résultats de performances s'appuient sur des tests à la date telle que décrit dans les configurations et peuvent ne pas refléter la totalité des mises à jour disponibles publiquement. Pour obtenir plus de détails, veuillez lire les informations de configuration. Aucun produit ou composant ne peut être totalement sécurisé en toutes circonstances.

Vos coûts et résultats peuvent varier.

Intel ne contrôle ni n'audite les données de parties tierces. Nous vous recommandons de consulter d'autres sources afin de confirmer si les données référencées sont exactes.

Les technologies Intel® peuvent nécessiter du matériel, des logiciels ou l'activation de services compatibles.

© Intel Corporation. Intel, le logo Intel et les autres marques Intel sont des marques commerciales d'Intel Corporation ou de ses filiales.
Les autres noms et marques peuvent être revendiqués comme la propriété de tiers.

0623/MH/MESH/353917-001US