Fiche produit

Intel® AI Engines
Processeurs Intel® Xeon® Scalable de 5e génération

Avec les processeurs Intel® Xeon® Scalable de 5e génération et les Intel® AI Engines, boostez la performance du pipeline d'IA

65 %
de l'inférence de l'IA dans les centres de données **s'appuie sur les processeurs Intel® Xeon®**[1]

Jusqu'à
14 fois plus
de performance d'inférence pour la détection d'objets en temps réel (SSD-ResNet34) avec les processeurs Intel® Xeon® Scalable de 5e génération dotés d'AMX BF16 par rapport aux processeurs Intel® Xeon® de 3e génération[2]

Jusqu'à
9,9 fois plus
de performance d'inférence pour le traitement du langage naturel (BERT-large) et 7,7 fois plus de performance/watt avec les processeurs Intel® Xeon® Scalable de 5e génération dotés d'AMX BF16 par rapport aux processeurs Intel® Xeon® de 3e génération[3]

Jusqu'à
8,7 fois plus
de performance d'inférence pour les systèmes de recommandation par lots (DLRM) et 6,2 fois plus de performance/watt avec les processeurs Intel® Xeon® de 5e génération par rapport aux processeurs Intel® Xeon® de 3e génération[4]

L'IA couvre un large éventail de charges de travail et de cas d'utilisation : prétraitement des données, apprentissage automatique classique (Machine Learning) ou applications d'apprentissage profond tels que le traitement du langage naturel et la reconnaissance d'images. Les processeurs Intel® Xeon® Scalable améliorent les performances de traitement globales du pipeline d'IA. Ils possèdent des accélérateurs intégrés optimisés pour des charges de travail d'IA spécifiques portant sur le Machine Learning, l'analyse des données et l'apprentissage profond.

Une puissance intégrée pour l'IA partout dans l'entreprise

L'IA se généralise, s'étendant à des charges de travail diverses et critiques. Le Machine Learning classique et l'apprentissage profond deviennent des éléments constitutifs du fonctionnement des entreprises, depuis les applications métiers jusqu'aux assistants vocaux automatisés. Cette généralisation de l'IA nécessite un long pipeline de développement, du prétraitement des données à la formation et au déploiement. Chaque étape s'accompagne de chaînes d'outils de développement, de frameworks et de charges de travail précis qui créent chacun des goulets d'étranglement uniques et posent des contraintes distinctes aux ressources informatiques. Dotés d'accélérateurs intégrés, les processeurs Intel® Xeon® Scalable sont prêts à exécuter les instructions dans l'intégralité du pipeline, augmentant les performances de l'IA à tous les niveaux.

Intel® Accelerator Engines : des accélérateurs intégrés spécialement conçus pour prendre en charge les charges de travail émergentes les plus exigeantes

Les processeurs Intel® Xeon® Scalable de 5e génération excellent dans le calcul général et continueront de sous-tendre de nombreuses charges de travail critiques d'IA existantes. Ces processeurs sont dotés d'Intel® AMX (Intel® Advanced Matrix Extensions), un accélérateur d'IA intégré conçu pour accélérer l'inférence et l'entraînement par apprentissage profond sur le processeur. Dans de nombreux cas, les surcoûts et la complexité liés à un accélérateur discret parviennent à être éliminés. La dernière génération de processeurs Intel® Xeon® convient bien aux grands modèles de langage (LLM) sollicitant moins de 20 milliards de paramètres, ce qui satisfait généralement aux accords de niveau de service des clients[5]. Intel® AMX se distingue également dans l'apprentissage par transfert et le réglage fin. L'entraînement d'un modèle ne s'étend plus sur des journées ou des heures. Forts des processeurs Intel® Xeon® assurant 65 % d'inférence dans les centres de données, les clients pourront exploiter leur architecture existante pour l'IA à usage général, sans s'égarer dans les arcanes d'un passage à une infrastructure GPU.

Avec les processeurs Intel® Xeon® Scalable de 5e génération et les moteurs Intel® Accelerator, les innovations de demain prennent vie

Vous utilisez des processeurs Intel® Xeon® pour vos charges de travail sur site, dans le Cloud ou à la périphérie ? Alors, les processeurs Intel® Xeon® dotés d'Intel® Accelerator Engines intégrés peuvent stimuler votre activité, grâce aux atouts qu'ils offrent, notamment une protection des données accrue et une optimisation de l'infrastructure.

Témoignage client : l'accélération avec les processeurs Intel® Xeon® Scalable en prise avec la réalité d'entreprise

Tencent Cloud propose une synthèse vocale en temps réel grâce aux processeurs Intel® Xeon® Scalable.

Tout savoir de ce projet ›

Gunpowder exécute des instances Google Cloud C3 à l'aide de CPU Intel® Xeon® de 4e génération pour accélérer les performances de rendu.

Lire le témoignage ›

Les Intel® Accelerator Engines peuvent également servir à exploiter plus largement les CPU virtuels et physiques et à réduire le nombre de licences de solutions par cœur. Ils permettent avant tout d'accroître les performances des applications, de réduire les coûts et d'améliorer l'efficacité des plateformes.

Accélérez l'apprentissage profond avec Intel® Advanced Matrix Extensions

Intel® AMX constitue la dernière avancée d'Intel en matière d'entraînement et d'inférence par apprentissage profond concernant les processeurs Intel® Xeon® Scalable de 5e génération. Conçu pour le traitement du langage naturel, les systèmes de recommandation, la reconnaissance d'images entre autres charges de travail, Intel® AMX aide les clients à atteindre des performances d'inférence de classification d'objets en temps réel jusqu'à 7,2 fois plus élevées avec des performances/watt 5,3 fois plus élevées sur les processeurs Intel® Xeon® de 5e génération dotés d'AMX BF16 par rapport aux processeurs Intel® Xeon® de 3e génération[6].

Avec Intel® AMX, comme la charge de travail des modèles d'IA est amplifiée, les clients sont plus nombreux à pouvoir tenir leurs accords de niveau de service sur les plateformes qu'ils exploitent déjà. Les processeurs Intel® Xeon® Scalable de 5e génération offrent des fréquences turbo améliorées pour les charges de travail possédant une affinité avec les opérations vectorielles et matricielles, notamment le calcul à haute performance et l'IA, avec l'ajout de cinq niveaux de ratios turbo.

Intel® AMX améliore les performances des opérations de multiplication matricielle avec un débit plus élevé (Ops/Cycle) par rapport à Intel® Advanced Vector Extensions 512 (Intel® AVX-512) sur les cœurs de CPU[7]. Les charges de travail de formation en apprentissage profond peuvent être ainsi achevées plus rapidement et les clients sont plus nombreux à pouvoir tenir leurs accords de niveau de service sur les plateformes qui sous-tendent leur activité.

Des technologies en appui du traitement du langage naturel et de l'IA générative

Les processeurs Intel® Xeon® Scalable de 5e génération dotés d'Intel® AMX offrent des performances nettement accrues en matière de traitement du langage naturel, sans matériel supplémentaire. Les bibliothèques Intel sont optimisées pour TensorFlow et PyTorch auxquelles elles sont également intégrées, afin que les développeurs bénéficient d'une accélération intégrée de l'IA. Ces derniers peuvent également procéder plus facilement à la migration du code de différents environnements matériels, un processus qui peut parfois s'avérer fastidieux et onéreux.

En accélérant l'inférence et l'entraînement par apprentissage profond, le processeur Intel® Xeon® Scalable de 5e génération doté d'Intel® AMX vous aide à tenir vos accords de niveau de service pour un coût total de possession raisonnable, grâce à un système de recommandation basé sur l'apprentissage profond tenant compte des signaux de comportement de l'utilisateur en temps réel et de caractéristiques contextuelles supplémentaires (horaire, lieu…).

Le processeur de 5e génération traite également des modèles d'IA générative qui reproduisent un contenu axé sur le facteur humain, en prenant en charge de grands modèles de langage et la génération d'images à partir de texte. Pour les tâches d'IA générative plus intensives, il est possible d'utiliser l'accélérateur d'IA Intel® Gaudi®, le GPU Intel® Data Center et d'autres composants matériels pour élargir les capacités du processeur.

Intel® AVX-512 : pour un Machine Learning plus performant

Les processeurs Intel® Xeon® peuvent hacher le chiffrement SSL de sites web, analyser des bases de données considérables, voire procéder à des simulations en recherche pharmaceutique, en conception de puces ou de moteurs de voitures de course.

Au fil des générations, les améliorations apportées à Intel® Advanced Vector Extensions 512 (Intel® AVX-512) permettent aux processeurs Intel® Xeon® Scalable d'intégrer plus d'opérations dans chaque cycle d'horloge et d'accroître les performances des applications de traitement parallèle. L'architecture du jeu d'instructions (ISA) Intel® AVX-512 comprend des extensions conçues pour stimuler les performances de diverses charges de travail dans les domaines de l'IA, du HPC (calcul haute performance), de la mise en réseau et du stockage.

Les performances turbo passent de quatre à cinq niveaux de ratios turbo pour la nouvelle génération de processeurs, améliorant les fréquences turbo pour certaines charges de travail liées au HPC et à l'IA, grâce à l'appui d'Intel® AMX et Intel® AVX-512.

Moins d'étapes pour un traitement plus rapide

Les mathématiques peuvent allier intelligence et élégance. Sur les processeurs Intel® Xeon® Scalable de 5e génération, Intel® AVX-512 sollicite l'intelligence et la beauté mathématiques pour condenser, combiner et fusionner des opérations de calcul courantes en moins d'étapes. Voici un exemple rudimentaire : pour qu'une unité centrale calcule 3 x 3 x 3 x 3 x 3, cinq cycles d'horloge sont nécessaires. Si vous créez une instruction pour 35, l'unité centrale peut effectuer ce calcul en un cycle. Intel® AVX-512 suit cette logique et l'applique à des centaines d'opérations spécifiques à la charge de travail, notamment certains calculs des plus complexes en IA.

Compter par 8 plutôt que de tout compter à la suite

Le « 512 » dans Intel® AVX-512 renvoie à cette seconde stratégie d'instructions destinée à augmenter le nombre de bits à disposition du processeur à chaque cycle d'horloge. Il y a quarante ans, les PC 16 bits étaient assez impressionnants, mais les machines 32 bits les ont dépassés. Aujourd'hui, votre smartphone fonctionne en 64 bits. Le nombre de bits fait référence au nombre de registres (ces emplacements de mémoire où l'unité centrale stocke les données) que l'unité centrale peut traiter par cycle d'horloge. Comme son nom l'indique, Intel® AVX-512 porte le nombre de registres à 512 bits. Lorsqu'une application tire parti d'Intel® AVX-512, elle fonctionne jusqu'à huit fois plus vite que la vitesse de base de 64 bits du CPU, par la simple augmentation du nombre de registres. Un peu comme si on comptait jusqu'à 96 par 1, 2, 3... au lieu de 8, 16, 24.

Avec les processeurs Intel® Xeon®, l'accélération de l'IA est pratiquement automatique

Avec les processeurs Intel® Xeon® Scalable, l'accélération de l'IA est intégrée dans l'architecture de jeu d'instructions du processeur. En d'autres termes, le fonctionnement est immédiat pour tout logiciel qui peut en tirer parti. Les ingénieurs logiciels d'Intel® ne cessent d'optimiser les chaînes d'outils d'IA open-source, diffusant ces optimisations à la communauté des développeurs, comme en témoigne TensorFlow 2.9, livré par défaut avec les optimisations de la bibliothèque Intel® oneDNN (Intel® oneAPI Deep Neural Network Library). Téléchargez la dernière édition pour bénéficier automatiquement des optimisations d'Intel® dans TensorFlow.

D'autres applications du pipeline d'IA sont proposées en accès libre : les experts en mégadonnées et les développeurs peuvent ainsi télécharger des distributions, bibliothèques et environnements de développement Intel open-source tirant parti de chaque accélérateur intégré dans notre architecture de jeu d'instructions pour processeurs Intel® Xeon® Scalable. Pourquoi ces spécialistes devraient-ils recoder leurs outils et les recompiler pour Intel® AVX-512 alors qu'on peut le faire pour eux ?

Aujourd'hui, toute entreprise doit pouvoir tirer de son infrastructure des performances accrues pour les charges de travail à moindre coût. Grâce à la conception dédiée des moteurs Intel AI intégrés aux processeurs Intel® Xeon® Scalable, vous pourrez tirer le meilleur parti des charges de travail d'IA indispensables à votre activité.

Découvrez ce que les processeurs Intel® Xeon® Scalable dotés des Intel® Accelerator Engines intégrés peuvent accomplir pour les charges de travail d'IA indispensables à votre activité !

En savoir plus

IA et apprentissage profond sur les processeurs Intel® Xeon® Scalable (AI and Deep Learning on Intel® Xeon® Scalable Processors) ›

Intel® AVX-512 ›

Boîte à outils Intel® AI Analytics ›

Développer sur du matériel et des logiciels Intel® (Developing on Intel® Hardware and Software) ›

Accélérez les charges d'IA dans le Cloud ou sur votre infrastructure grâce aux optimisations Intel pour l'IA et le Machine Learning.

En savoir plus ›

intel® XEON®

1. Basé sur la modélisation du marché Intel de la base installée mondiale de serveurs de centres de données exécutant des charges de travail d'inférence IA en décembre 2022.
2. Voir [A21] sur intel.com/processorclaims: processeurs Intel® Xeon® Scalable de 5e génération. Les résultats effectifs peuvent varier
3. Voir [A19] sur intel.com/processorclaims: processeurs Intel® Xeon® Scalable de 5e génération. Vos résultats peuvent varier.
4. Voir [A20] sur intel.com/processorclaims: processeurs Intel® Xeon® Scalable de 5e génération. Vos résultats peuvent varier.
5. Basé sur une modélisation interne d'Intel en décembre 2023.
6. Voir [A22] sur intel.com/processorclaims : processeurs Intel® Xeon® Scalable de 5e génération. Vos résultats peuvent varier.
7. https://edc.intel.com/content/www/us/en/products/performance/benchmarks/vision-2022/, Test de performance : sessions #41 et #42. Vos résultats peuvent varier.