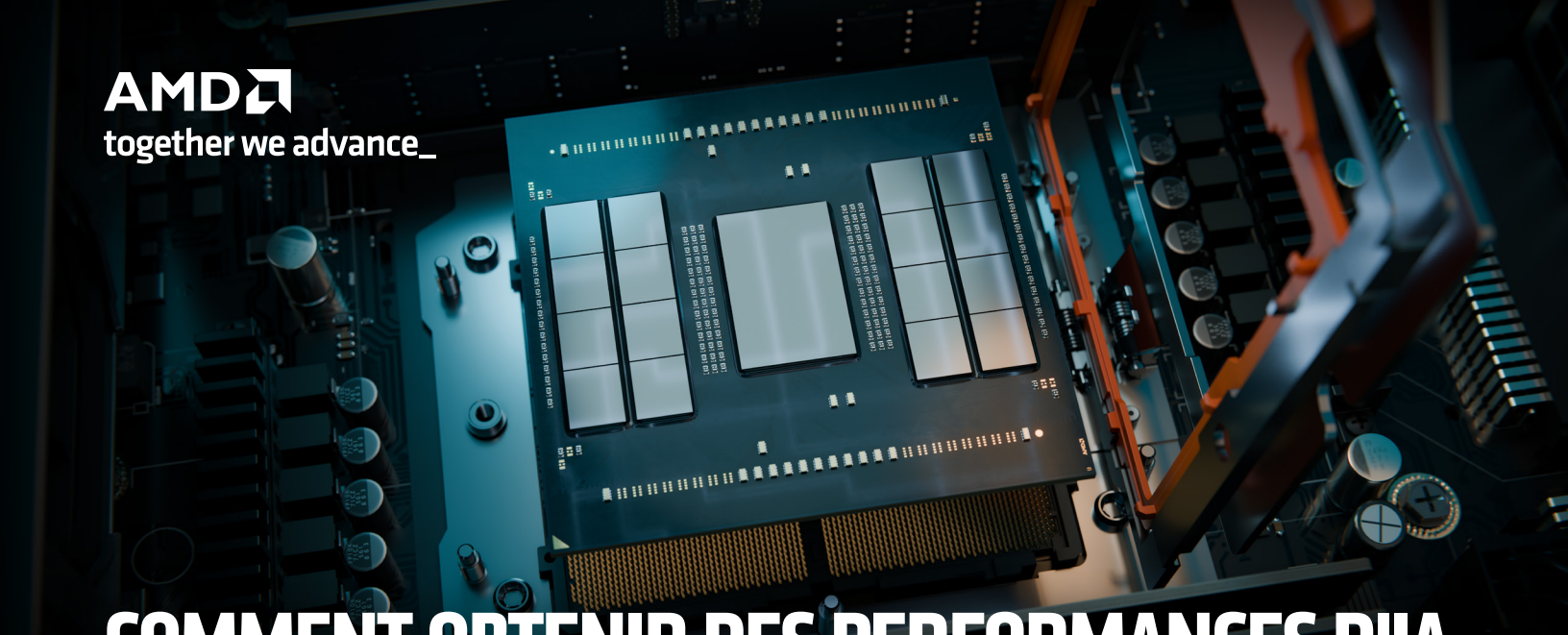
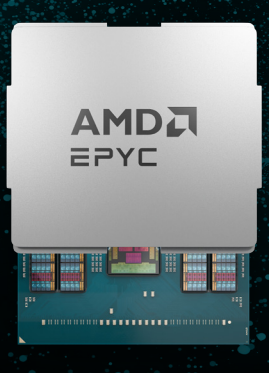
AMD
together we advance_



COMMENT OBTENIR DES PERFORMANCES D'IA *JUSQU'À 20 % SUPÉRIEURES* AVEC DES GPU HAUTES PERFORMANCES[1,2]



Les CPU AMD EPYC™ de 5e génération comprennent des modèles à plus haute fréquence conçus spécifiquement pour l'hébergement de plateformes pour accélérateurs. Ces CPU excellent dans l'orchestration des mouvements de données et la gestion de plusieurs machines virtuelles, des capacités critiques qui permettent d'extraire des performances hors pair des plateformes GPU[1,2].

Le bon hôte fait la différence

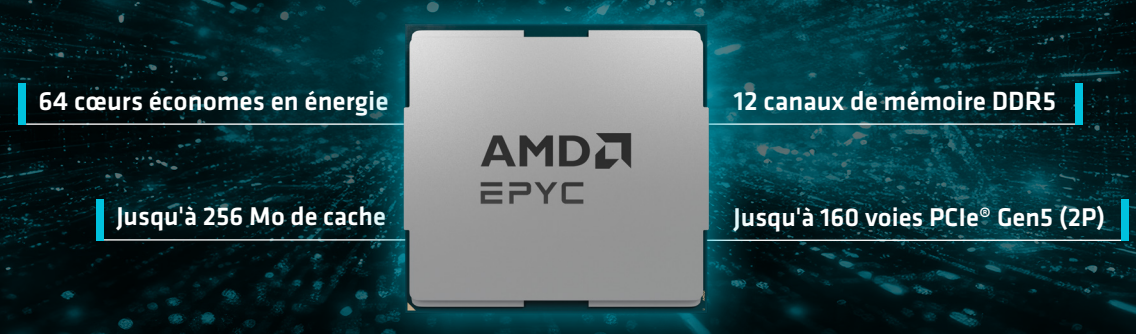
LA SÉLECTION DE CPU DE NŒUD D'HÔTE AMD EPYC ACCROÎT LE DÉBIT D'INFÉRENCE ET D'ENTRAÎNEMENT

CPU HÔTE	PLATEFORME GPU	INFÉRENCE	ENTRAÎNEMENT
CPU 2P AMD EPYC™ 9575F (128 cœurs au total)	8 GPU AMD Instinct™ MI300X	Llama 3.1-70B @ FP8 **Jusqu'à 8 %** de jetons/s en plus[3] 700 000 jetons/seconde supplémentaires sur un cluster de 1 000 nœuds de GPU AMD Instinct	Ensemble d'entraînement DeepSpeed 0.14.0 @ FP8 Stable Diffusion XL v2 **Jusqu'à 20 %** d'échantillons/s en plus[4]
CPU HÔTE	PLATEFORME GPU	INFÉRENCE	ENTRAÎNEMENT
CPU 2P AMD EPYC™ 9575F (128 cœurs au total)	8 GPU NVIDIA H100	Llama 3.1-70B @ FP8 **Jusqu'à 20 %** de jetons/s en plus[5]	Llama 3.1-8B @ BF16 Longueur de séquence max. 1 024 **Jusqu'à 15 %** d'échantillons/s en plus[6]

Performances comparées à 2P Intel® Xeon® Platinum 8592+ (128 cœurs au total) hébergeant les mêmes GPU et exécutant des charges de travail identiques.

Conçu pour booster les performances des accélérateurs d'IA

L'EPYC 9575F à 5 GHz de boost max est 28 % plus performant que l'Intel® Xeon® Platinum 8592+[7].



Une gamme d'options pour héberger des GPU

Avec des fréquences allant jusqu'à 5 GHz et la prise en charge de jusqu'à 6 To de mémoire, les CPU AMD EPYC de 5e génération offrent plusieurs modèles spécialement conçus pour les clusters de GPU.

PROCESSEUR	NOMBRE DE CŒURS	FRÉQUENCE BOOST MAX
9575F	64	5 GHz
9475F	48	4,8 GHz
9375F	32	4,8 GHz
9275F	24	4,8 GHz
9175F	16	5 GHz

CPU AMD EPYC™ DE 5e GÉNÉRATION : *LE MEILLEUR CPU POUR L'IA EN ENTREPRISE*[8]

Bénéficiez de performances serveur de pointe pour les charges de travail d'IA, d'entreprise et de cloud avec les processeurs AMD EPYC de 5e génération.

Découvrez AMD EPYC

1. Résultats de l'entraînement Stable Diffusion XL v2 basé sur des tests internes AMD réalisés le 10 octobre 2024. Configurations SDXL : DeepSpeed 0.14.0, TP8 parallèle, FP8, lot de 24, résultats en secondes 2P AMD EPYC 9575F (128 cœurs au total) avec 8 AMD Instinct MI300X-NPS1-SPX-192GB-750W, interconnectivité GPU XGMI, ROCm™ 6.2.0-66, 2 304 Go 24x96 Go DDR5-6000, BIOS 1.0 (déterminisme de la puissance = désactivé), Ubuntu® 22.04.4 LTS, noyau 5.15.0-72-generic, 334.80 secondes 2P Intel Xeon Platinum 8592+ (128 cœurs au total) avec 8 AMD Instinct MI300X-NPS1-SPX-192GB-750, interconnectivité GPU XGMI, ROCm 6.2.0-66, 2 048 Go 32x64 Go de DDR5-4400, BIOS 2.0.4 (déterminisme de la puissance = désactivé), Ubuntu 22.04.4 LTS, noyau 5.15.0-72-generic, 400,43 secondes, soit une augmentation de 19,600 % des performances d'entraînement. Les résultats varient selon plusieurs facteurs, tels que les configurations système, les versions logicielles et les paramètres du BIOS. (9xx5-059A)
2. Résultats du débit d'inférences Llama3.1-70B basés sur les tests internes d'AMD en date du 01/09/2024. Configurations Llama3.1-70B : TensorRT-LLM 0.9.0, nvidia/cuda 12.5.0-devel-ubuntu22.04, FP8, configurations de jeton d'entrée/de sortie (cas d'utilisation) : [BS=1 024 E/S=128/128, BS=1 024 E/S=128/2 048, BS=96 E/S=2 048/128, BS=64 E/S=2 048/2 048]. Résultats en jetons/seconde. 2P AMD EPYC 9575F (128 cœurs au total) avec 8 NVIDIA H100 80 Go HBM3, 1,5 To 24x64 Go de DDR5-6000, NVMe® Micron_9300_MTFDHAL3T8T8TDP 3 To 1 Gbit/s, BIOS T20240805173113 (déterminisme=puissance, SR-IOV=activé), Ubuntu 22.04.3 LTS, noyau=5.15.0-117-generic (atténuations=désactivées, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/sys/vm/drop_caches), 2P Intel Xeon Platinum 8592+ (128 cœurs au total) avec 8 NVIDIA H100 80 Go HBM3, 1 To 16x64 Go de DDR5-5600, NVMe® Dell Ent PPM1735a MU 3,2 To, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic (processor.max_cstate=1, intel_idle.max_cstate=0 atténuations=désactivées, cpupower frequency-set -g performance), BIOS 2.1 (performances maximales, SR-IOV=activé), taille du lot de jetons d'entrée/de sortie EMR Turin relative 128/128 1024 814.678 1101.966 1.353 128/2048 1024 2120.664 2331.776 1.1 2048/128 96 114.954 146.187 1.272 2048/2048 64 333.325 354.208 1.063 pour une multiplication moyenne du débit de 1,197. Les résultats varient selon plusieurs facteurs, tels que les configurations système, les versions logicielles et les paramètres du BIOS. (9xx5-014)
3. Au 10/10/2024, ce scénario contient plusieurs hypothèses et estimations. Bien que basé sur les recherches internes d'AMD et sur les meilleures approximations, il doit être considéré comme un exemple fourni uniquement à titre indicatif et ne saurait se substituer à des tests réels lors de prises de décision. Référence 9xx5-056A : « Serveur propulsé par 2P AMD EPYC 9575F et 8 GPU AMD Instinct MI300X exécutant des charges de travail d'inférence sélectionnées Llama3.1-70B avec une précision FP8 par rapport à un serveur propulsé par 2P Intel Xeon Platinum 8592+ et 8 GPU AMD Instinct MI300X, avec une augmentation globale de débit de ~8 % à travers des cas d'utilisation d'inférence sélectionnés » et 8 763,52 jetons/s (9575F) par rapport 8 048,48 jetons/s (8592+) à 128 jetons d'entrée / 2 048 jetons de sortie, 500 invites pour 1,089 x les jetons/s ou 715,04 jetons/s de plus. 1 nœud = 2 CPU et 8 GPU. En supposant un cluster de 1 000 nœuds, 1 000 x 715,04 = 715 040 jetons/s pour ~700 000 jetons/s supplémentaires. Les résultats varient selon plusieurs facteurs, tels que les configurations système, les versions logicielles et les paramètres du BIOS. (9xx5-087)
4. Voir remarque 1 ci-dessus.
5. Voir remarque 2 ci-dessus.
6. résultats des tests d'entraînement Llama3.1-8B (BF16, longueur de séquence max. 1 024) basés sur les tests internes d'AMD en date du 05/09/2024. Configurations Llama3.1-8B : longueur de séquence max. 1 024, BF16, docker : huggingface/transformers-pytorch-gpu:latest 2P AMD EPYC 9575F (128 cœurs au total) avec 8 NVIDIA H100 80 Go HBM3, 1,5 To 24x64 Go de DDR5-6000, NVMe® Micron_9300_MTFDHAL3T8T8TDP 3 To 1 Gbit/s, BIOS T20240805173113 (déterminisme=puissance, SR-IOV=activé), Ubuntu 22.04.3 LTS, noyau=5.15.0-117-generic (atténuations=désactivées, cpupower-frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/sys/vm/drop_caches), pour 31,79 échantillons d'entraînement/seconde. 2P Intel Xeon Platinum 8592+ (128 cœurs au total) avec 8 NVIDIA H100 80 Go HBM3, 1 To 16x64 Go de DDR5-5600, NVMe® Dell Ent PM1735a MU 3,2 To, Ubuntu 22.04.3 LTS, noyau 5.15.0-118-generic, (processor.max_cstate=1, intel_idle.max_cstate=0 atténuations=désactivé, cpupower frequency-set -g performance), BIOS 2.1, (performances maximales, SR-IOV=activé), pour 27,74 échantillons d'entraînement/seconde, soit une multiplication moyenne du débit de 1,146. Les résultats varient selon plusieurs facteurs, tels que les configurations système, les versions logicielles et les paramètres du BIOS. (9xx5-015)
7. Comparaison entre le CPU AMD EPYC 9575F de 5e génération à la fréquence la plus élevée (jusqu'à 5 GHz) et le CPU Intel Xeon Platinum 8592+ à la fréquence la plus élevée (jusqu'à 3,9 GHz), sur la base des spécifications publiées.
8. Comparaison basée sur la densité des threads, les performances, les fonctionnalités, la technologie de processus et les fonctions de sécurité intégrées au 10/10/2024. Les CPU de la série EPYC 9005 offrent la plus forte densité de threads, sont à la pointe du secteur avec plus de 500 records mondiaux de performance, y compris les records mondiaux de performance en termes d'opérations Java®/s en entreprise, dominent le secteur HPC en termes de performances de calcul de débit en virgule flottante, offrent des performances TPCx-AI d'IA de bout en bout et obtiennent les scores d'efficacité énergétique les plus élevés. La série EPYC de 5e génération offre des canaux de mémoire DDR5 en plus, avec plus de bande passante mémoire, prend en charge 70 % de voies PCIe® Gen5 en plus pour le débit d'E/S, offre jusqu'à 5 fois plus de cache L3 par cœur pour un accès plus rapide aux données. La série EPYC 9005 utilise la technologie avancée 3-4Nm et offre des fonctions de sécurité Secure Memory Encryption + Secure Encrypted Virtualization (SEV) + SEV Encrypted State + SEV-Secure Nested Paging. (EPYC-029D)

AMD
together we advance_