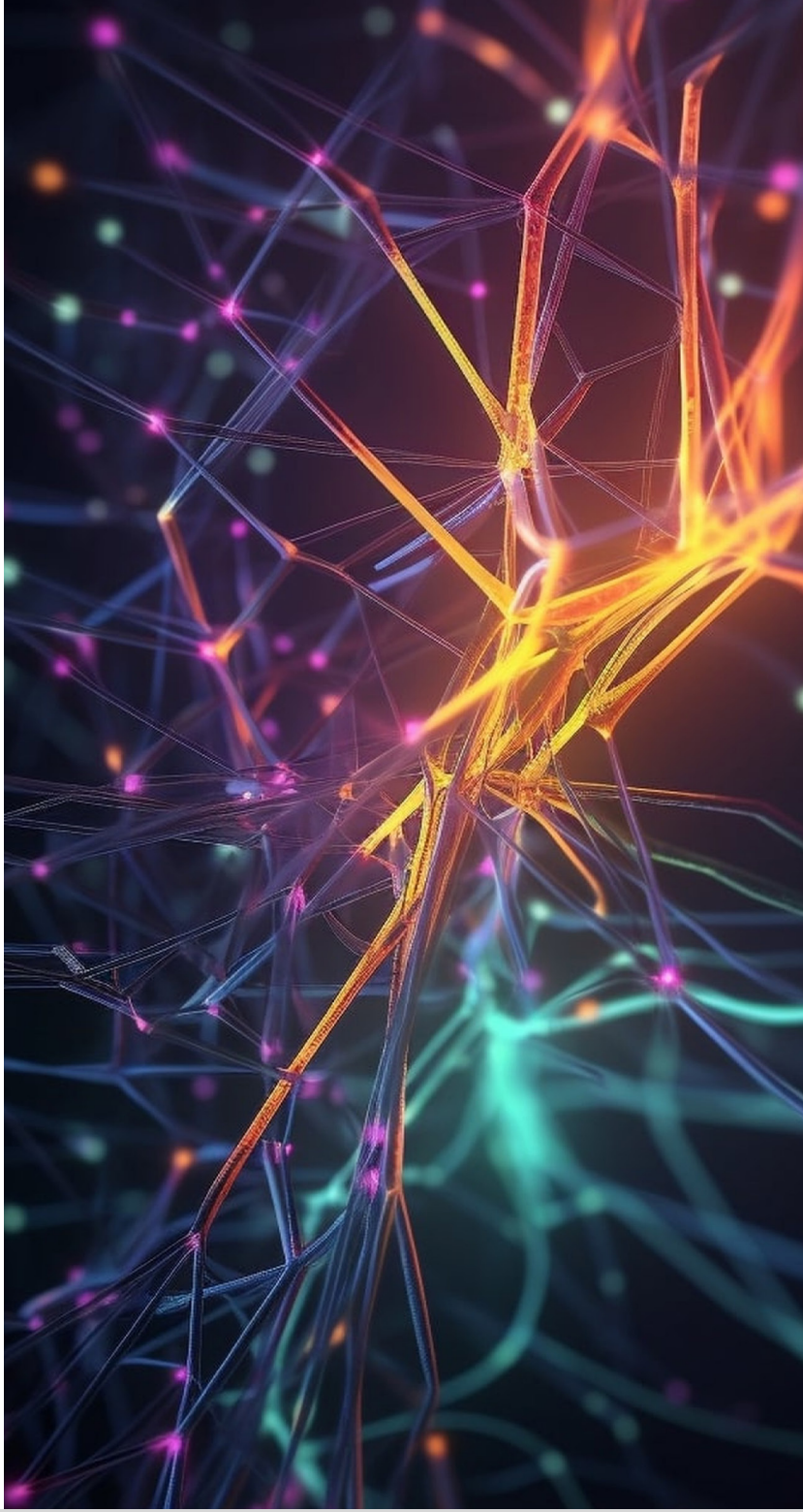

PRÉPARER LES
ENTREPRISES
À L'ÈRE DE L'IA
AMD
together we advance_

TABLE DES MATIÈRES



INTRODUCTION

L'IA : UN IMPÉRATIF POUR LES ENTREPRISES

L'intelligence artificielle connaît une expansion exceptionnelle qui a rarement été observée dans l'histoire des technologies professionnelles. Ainsi, l'IA est devenue un impératif pour les entreprises, car chaque fonction clé peut en bénéficier pour prendre des décisions plus rapidement, pour accroître la productivité et pour réduire les coûts opérationnels.

Cependant, cet essor rapide de l'IA ne se fait ni simplement ni sans heurts au sein des organisations. Leurs défis vont de la prise en compte de nouvelles exigences en matière de calcul et de consommation énergétique des centres de données, à l'alignement des investissements dans des ressources sur leurs objectifs (gains de productivité, accélération de l'innovation, etc.).

Cet e-book explique comment relever ces défis. Il contient des analyses pour guider les entreprises face à la nécessaire modernisation de leur infrastructure, pour réduire les risques d'erreurs coûteuses, d'opportunités d'innovation manquées et de hausse des coûts d'exploitation et de maintenance. Il leur montre également la voie vers une accélération des résultats commerciaux et un succès durable.

3 DÉFIS CLÉS POUR LES ENTREPRISES QUI ADOPTENT L'IA

1 LES INFRASTRUCTURES EXISTANTES NE SONT PAS PRÊTES POUR L'IA

Les limites des infrastructures existantes représentent le principal obstacle à l'adoption et à l'évolutivité de l'IA dans les entreprises.

Il est impératif que les entreprises intègrent des charges de travail d'IA, rapidement et à moindre coût. Or, ces charges de travail nécessitent des ressources de calcul importantes. On ne peut pas simplement les superposer sur des infrastructures de centres de données déjà saturées, dont la consommation énergétique est une préoccupation constante et dans lesquelles les ressources de CPU, le stockage et la bande passante réseau fonctionnent à pleine capacité ou presque.

L'ajout de racks supplémentaires de hardware compatible avec l'IA nécessite généralement plus de densité et d'espace au sol que ce que les centres de données existants peuvent offrir. De plus, les mises à niveau nécessaires pour augmenter la puissance ou la capacité de refroidissement peuvent s'avérer trop coûteuses. Les risques liés à la migration d'applications stratégiques existantes compliquent encore la modernisation, parallèlement à la création d'un nouvel environnement prêt pour l'IA

Enfin, les infrastructures existantes engendrent des coûts d'opportunité élevés : les ressources consacrées à leur maintenance ne peuvent pas être réaffectées à l'innovation en matière d'IA.

2 L'IA ACCROÎT LA DEMANDE EN MATIÈRE DE CALCUL CONFIDENTIEL

Alors que l'IA devient une application incontournable qui touche toutes les dimensions des données personnelles et professionnelles, la confiance dans la confidentialité et l'intégrité des données devient essentielle.

L'IA hautes performances repose souvent sur un hardware hétérogène (CPU, GPU et accélérateurs spécialisés), réparti sur plusieurs nœuds, voire sur plusieurs sites. Garantir une sécurité de bout en bout sur tous les appareils et toutes les liaisons réseau est un défi majeur. Même si les données sont chiffrées au repos, des vulnérabilités dans les environnements virtualisés ou conteneurisés peuvent permettre à des hyperviseurs malveillants d'accéder à des informations sensibles.

Les entreprises doivent donc intégrer des fonctionnalités de calcul confidentiel, comme Secure Encrypted Virtualization (SEV), dans leurs flux de travail d'IA. Et ces protections doivent pouvoir s'étendre à des dizaines, voire des centaines de nœuds de GPU/accélérateurs, en combinaison avec d'autres techniques de sécurité des données, pour créer un modèle fiable d'exécution de l'IA.

3 IL EST DIFFICILE D'INVESTIR DE MANIÈRE DÉCISIVE TOUT EN RESTANT FLEXIBLE

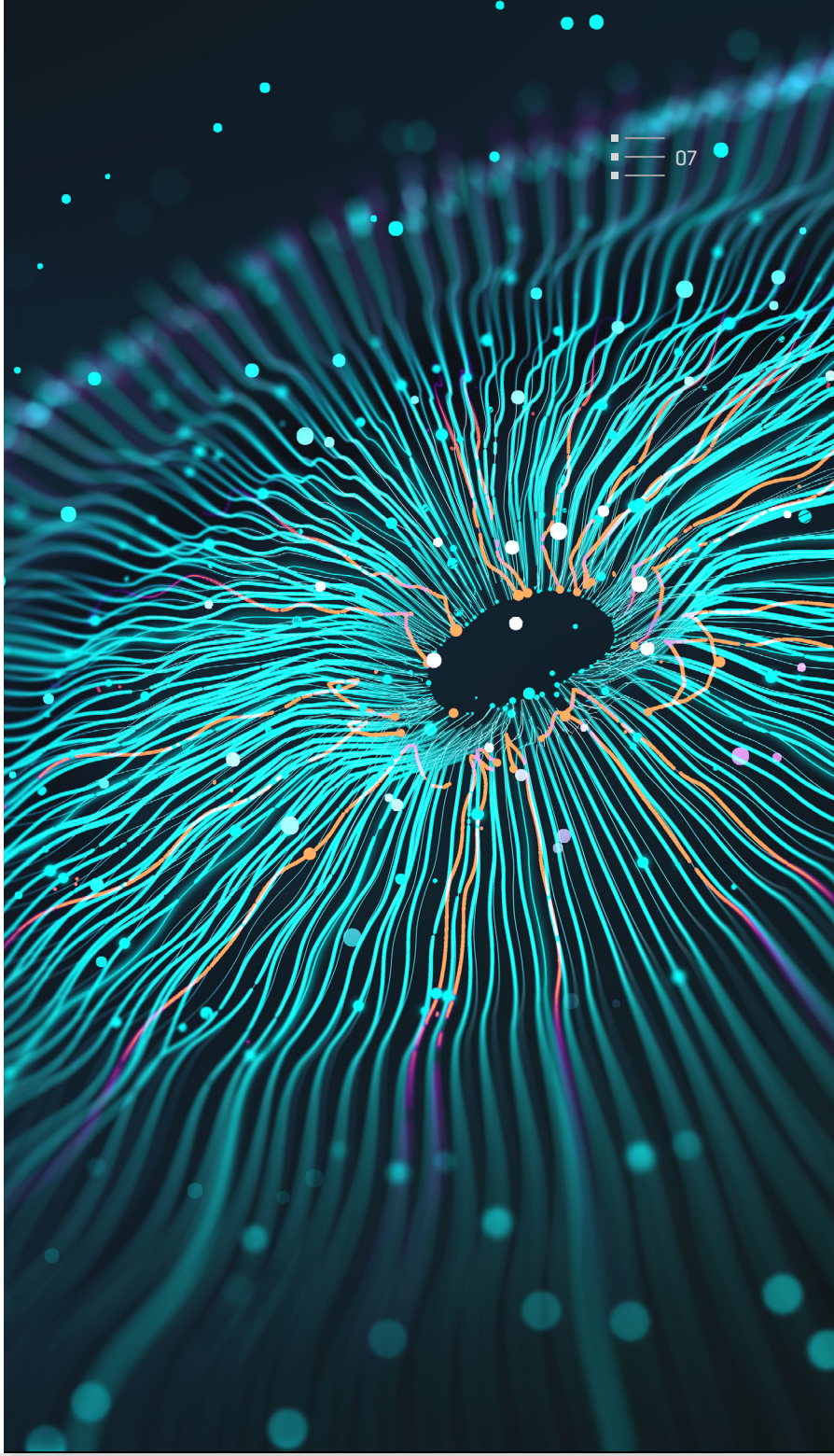
Les initiatives autour de l'IA poursuivent généralement deux grands objectifs : l'amélioration de la productivité ou l'innovation. Chacun impose des exigences différentes en matière d'infrastructure de centre de données.

Les initiatives d'IA axées sur l'augmentation de la productivité par l'automatisation ou l'optimisation des flux de travail relèvent souvent des charges de travail d'inférence, avec des besoins moindres en entraînement. Elles peuvent souvent être prises en charge efficacement par la dernière génération de CPU. En revanche, les projets visant de nouvelles sources de revenus ou des découvertes majeures ont des besoins en entraînement plus conséquents, nécessitant souvent des investissements dans de nouvelles ressources GPU.

Trouver le bon équilibre entre ces deux types d'investissements en IA est un défi stratégique majeur. Les avancées des frameworks et du hardware d'IA sont rapides, et chaque nouvelle génération de hardware ou de framework d'apprentissage automatique oblige à réévaluer les rapports coûts/performances.

DOMAINES CLÉS DE MODERNISATION AVEC L'IA AU NIVEAU DE L'INFRASTRUCTURE

L'accélération des résultats commerciaux grâce à l'IA implique de relever chacun de ces défis à l'aide d'une série de modernisations ciblées des centres de données.



DOMAINES CLÉS DE MODERNISATION AVEC L'IA AU NIVEAU DE L'INFRASTRUCTURE

OUVREZ LA VOIE À L'EXPANSION DE L'IA AVEC LA CONSOLIDATION DES SERVEURS

La consolidation des serveurs permet de réaliser des économies d'espace et d'énergie dans les centres de données, tout en ouvrant la voie à l'expansion de l'IA.

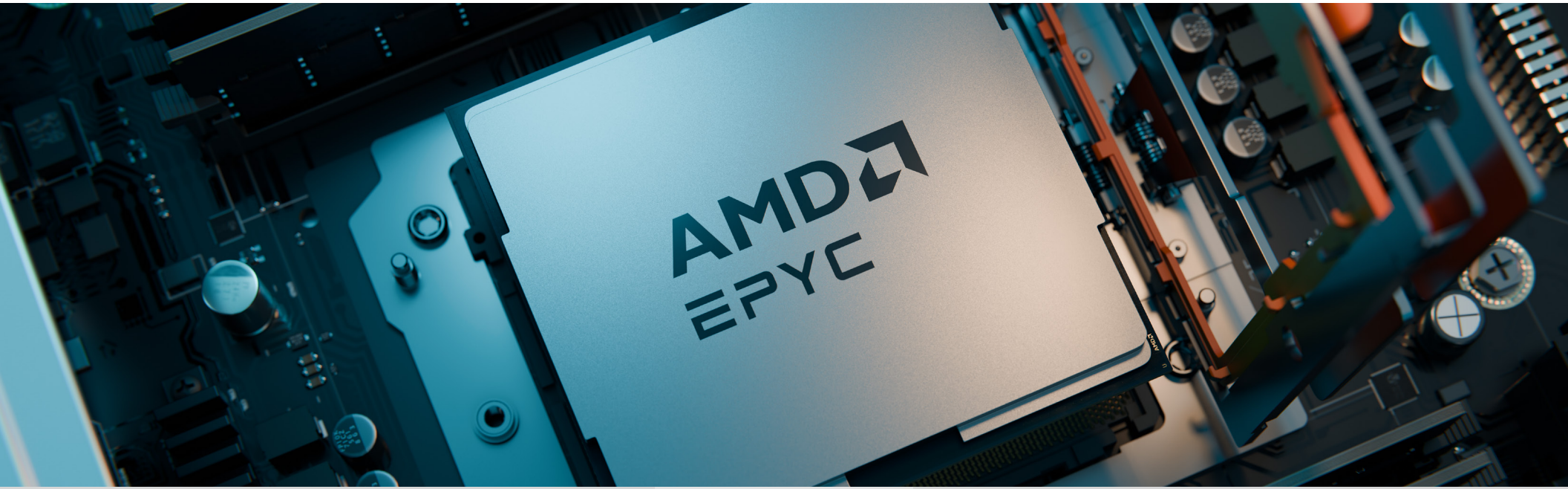
Les processeurs modernes offrent désormais des dizaines, voire des centaines de cœurs par socket. Cela permet un parallélisme massif pour le prétraitement des données d'IA et les tâches d'inférence à petite ou moyenne échelle. De cette façon, un seul serveur peut remplacer jusqu'à sept machines exitstantes[1]. Cela libère de l'espace au sol pour accueillir de nouveaux racks dédiés à l'IA, ainsi que les infrastructures de refroidissement avancées nécessaires.

La nouvelle génération d'architectures de CPU offre également des performances par watt supérieures à celles des générations précédentes, ce qui réduit la consommation d'énergie et les coûts opérationnels permanents liés à l'exécution des charges de travail d'IA.

PROCESSEURS AMD EPYC™ : LE CPU DE RÉFÉRENCE POUR L'IA[2]

Les processeurs AMD EPYC™ Série 9005 peuvent égaler les performances de calcul d'entiers du hardware existant avec jusqu'à 86 % de racks en moins[3], offrant un débit d'apprentissage automatique jusqu'à 3 fois supérieur par rapport aux processeurs Intel Xeon 8592+ à 64 cœurs[4].

Les cryptages au niveau du hardware, tels que la technologie SEV (Secure Encrypted Virtualization), permettent également aux modèles et aux données de rester cryptés pendant l'entraînement ou l'inférence, de sorte que la confidentialité reste préservée même dans les environnements multi-locataires.



DOMAINES CLÉS DE MODERNISATION AVEC L'IA AU NIVEAU DE L'INFRASTRUCTURE

ADOPTEZ UNE INFRASTRUCTURE FLEXIBLE POUR FAIRE ÉVOLUER L'IA PLUS RAPIDEMENT

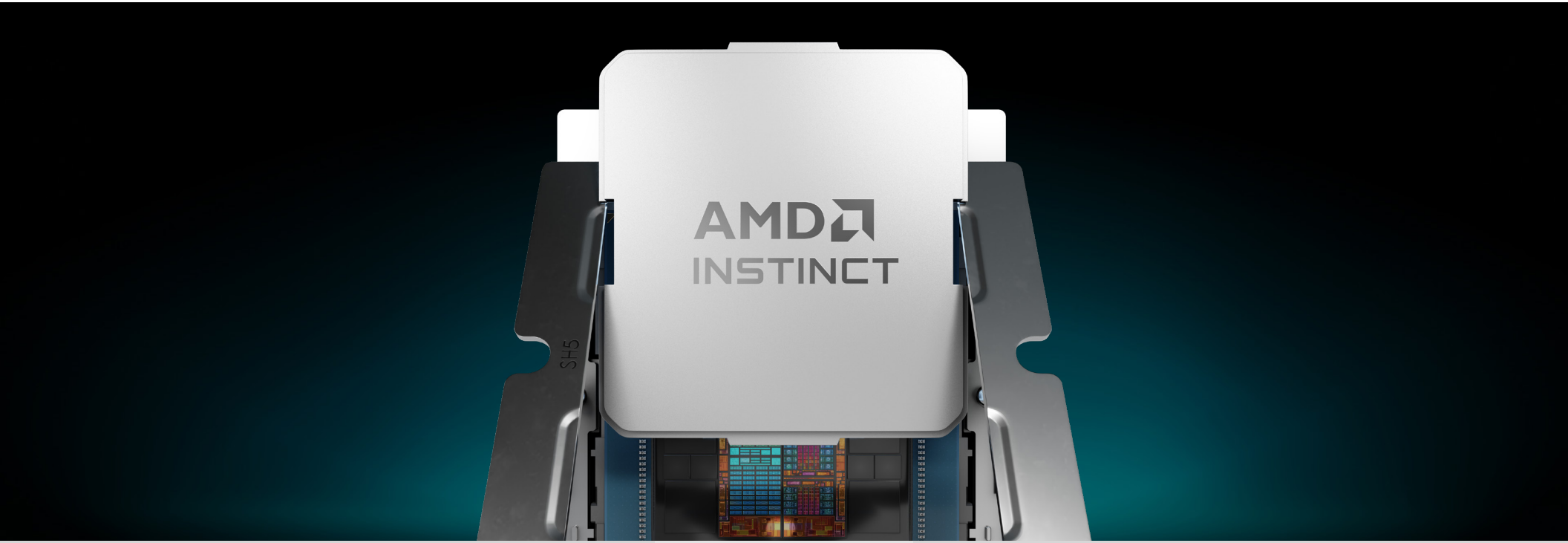
Des charges de travail d'IA diverses nécessitent différentes capacités de calcul. Lorsqu'une nouvelle infrastructure est requise, le choix de rester sur des architectures x86 plutôt que de passer à des options basées sur ARM permet aux entreprises de conserver leurs applications x86 existantes, ce qui simplifie les scale-out d'IA. Les entreprises peuvent également accélérer les délais de déploiement en tirant parti de bibliothèques, de conteneurs et d'implémentations de référence pré-optimisés, qui s'exécutent rapidement sur l'infrastructure de leur choix.

Le profil de calcul de cette infrastructure est ensuite défini par les besoins de l'entreprise concernée en matière d'IA. Le deep learning est très gourmand en données et nécessite une bande passante mémoire élevée ainsi qu'un traitement parallèle important, ce que les GPU offrent. À petite échelle, la plupart des tâches d'inférence peuvent être gérées efficacement par les CPU qui offrent des avantages en termes d'efficacité de calcul et de capacités d'orchestration des tâches. Combiner les forces du CPU avec la parallélisation intrinsèque du GPU permet de gérer les modèles les plus volumineux et de répondre à la croissance des besoins en IA.

Une infrastructure flexible et évolutive garantit non seulement des gains de performance à court terme, mais permet également aux entreprises de s'adapter aux futures avancées de l'IA, sans nécessiter une refonte technologique complète.

ACCÉLÉRATEURS AMD INSTINCT™ : DE HAUTES PERFORMANCES À N'IMPORTE QUELLE ÉCHELLE

Les accélérateurs AMD Instinct™ MI325X égalent les performances d'entraînement de 8 GPU concurrents[5], tout en atteignant des performances d'inférence jusqu'à 1,4 x supérieures[6]. Soutenus par une pile software d'IA en expansion rapide dans ROCm™, les accélérateurs AMD Instinct™ permettent une exécution fluide et immédiate de plus d'un million de modèles.



DOMAINES CLÉS DE MODERNISATION AVEC L'IA AU NIVEAU DE L'INFRASTRUCTURE

AMÉLIOREZ L'EFFICACITÉ DU RÉSEAU ET SIMPLIFIEZ LA GESTION

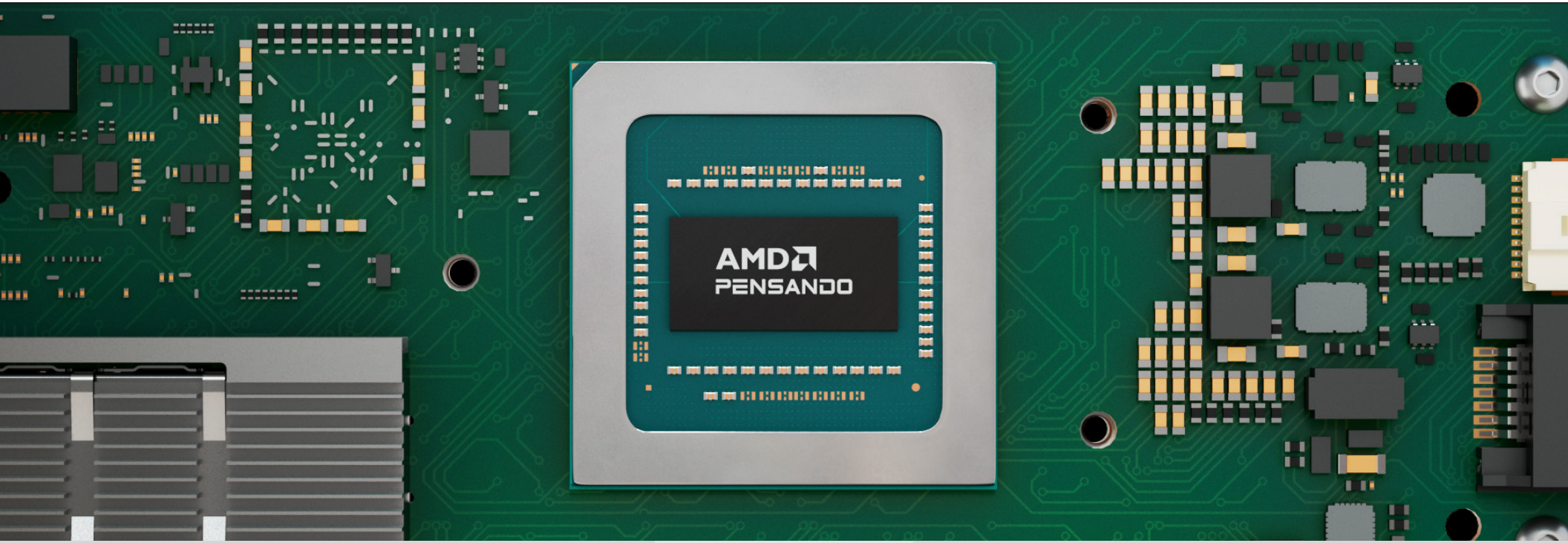
Le débit élevé et la sensibilité aux performances des charges de travail d'IA exercent une pression considérable sur les réseaux des centres de données.

Les cartes d'interface réseau (NIC) et les unités de traitement de données (DPU) sont des outils essentiels qui allègent le poids que les charges de travail d'IA exercent sur les réseaux dorsaux.

Les entreprises peuvent tirer parti des NIC intégrant des ressources de traitement supplémentaires pour délester le CPU hôte de certaines tâches. Ce délestage peut concerner notamment la classification des paquets, le cryptage/décryptage, l'accélération du plan de données et une gestion de la congestion plus avancée. En parallèle, les DPU dotées de cœurs programmables peuvent exécuter des services de mise en réseau, de sécurité ou de stockage, ce qui réduit encore la charge du CPU. Elles permettent également un déplacement direct des données entre le réseau, le stockage et les GPU/CPU, sans intervention importante de l'hôte.

MISE EN RÉSEAU AMD PENSANDO™ : AMÉLIORATION DES PERFORMANCES DANS LES RÉSEAUX D'IA

AMD Pensando™ Salina 400, une DPU entièrement programmable P4, fournit 2 fois plus de bande passante, de connexions par seconde, de paquets par seconde et d'opérations de stockage par rapport aux générations précédentes[7]. Par ailleurs, la NIC AMD Pensando Pollara 400 utilise la fonctionnalité RDMA compatible UEC, qui offre un temps d'exécution des messages 6 fois plus rapide et un temps de réalisation collectif 5 fois plus rapide que RoCEv2.



DOMAINES CLÉS DE MODERNISATION AVEC L'IA AU NIVEAU DE L'INFRASTRUCTURE

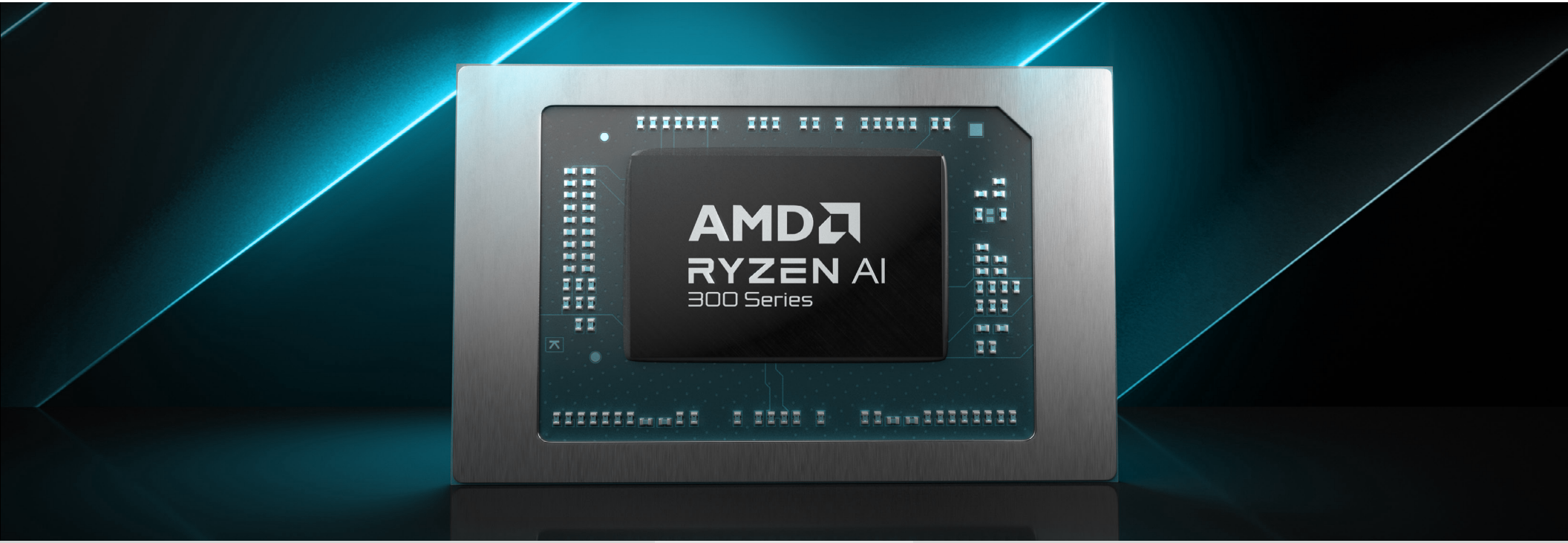
AU-DELÀ DU CENTRE DE DONNÉES

L'IA n'est pas réservée aux déploiements à grande échelle de serveurs dans le cloud ou en centre de données. Pour de nombreux utilisateurs finaux, les PC dotés de l'IA offrent de nombreux avantages au quotidien dans le cadre professionnel.

En traitant les données localement plutôt qu'en s'appuyant sur des ressources dans le cloud, ces PC IA peuvent fournir des performances en temps réel, même lorsque les connexions réseau ne sont pas optimales. Cette approche permet également de conserver des données personnelles ou sensibles sur l'appareil, renforçant ainsi la confidentialité. Les utilisateurs métiers bénéficient d'applications assistées par l'IA qui rationalisent les tâches de productivité, automatisent le traitement des données et fournissent des informations personnalisées basées sur les habitudes d'utilisation, le tout à une vitesse fulgurante.

PROCESSEURS AMD RYZEN™ : À LA POINTE DE L'ÈRE DE L'IA SUR PC

Les processeurs AMD Ryzen™ AI Série 300 prennent en charge des centaines d'expériences d'IA différentes et équipent une large gamme de PC portables Microsoft Copilot+ actuellement disponibles. Ces systèmes offrent des performances multithread jusqu'à 1,4 fois supérieures par rapport aux offres concurrentes[8], et jusqu'à 23 heures d'autonomie sur plusieurs jours[9].



CONNAÎTRE UN SUCCÈS DURABLE AVEC AMD

AMD : LE PARTENAIRE DES ENTREPRISES POUR SUIVRE L'ÉVOLUTION CONTINUE DE L'IA

En entreprise, l'IA s'intègre dans tous les domaines. Elle s'adapte à des tâches spécifiques, variant selon les cas d'utilisation et les services, et elle est de plus en plus spécialisée par secteur et par domaine.

Le portefeuille complet de solutions d'IA d'AMD offre à chaque entreprise un moyen d'accélérer l'adoption de l'IA et de réduire le temps nécessaire pour obtenir des résultats. Il permet également un contrôle précis des coûts opérationnels courants, grâce aux CPU AMD EPYC, reconnus pour leurs excellentes performances par watt. Les entreprises bénéficient donc d'une réduction de l'espace et de la consommation énergétique, ainsi que de coûts de licences plus faibles[10]. Elles s'assurent ainsi d'un succès durable.

AMD s'engage à maintenir un écosystème ouvert pour aider ses entreprises partenaires à suivre le rythme rapide de l'évolution de l'IA. Les clients d'AMD peuvent tirer parti des avancées de l'IA quelle que soit leur source, sans dépendance à un fournisseur. Cette approche ouverte de l'écosystème permet également une validation et une intégration simplifiées, une assistance immédiate de la part des partenaires clés et une conformité constante avec les réglementations et les meilleures pratiques du secteur.

PLANIFIEZ L'INTÉGRATION DE L'IA EN TOUTE CONFIANCE

La planification à long terme est essentielle pour la réussite de tout projet lié aux données en entreprise. S'associer à AMD, c'est collaborer avec un leader technologique stable, présent depuis des décennies, qui investit en permanence dans la recherche et le développement, et qui a une longue expérience dans la conception de feuilles de route de produits et d'objectifs de performances. Avec cette fiabilité, les entreprises sont tranquilles d'esprit et peuvent compter sur un partenaire technologique stable, dédié à leurs investissements et leurs plans en matière d'infrastructure.

Contactez-nous pour discuter de la manière dont AMD peut accélérer vos résultats commerciaux avec l'IA, et vous permettre d'atteindre une réussite évolutive et durable.

Échangeons

AFFIRMATIONS :

1. 9xx5TCO-002A : Ce scénario contient de nombreuses hypothèses et estimations et, bien que basé sur les recherches internes d'AMD et sur les meilleures approximations, il doit être considéré comme un exemple à titre informatif uniquement et non utilisé comme une base pour la prise de décision à la place de tests réels. L'outil d'estimation du TCO (coût total de possession) des émissions de gaz à effet de serre et des serveurs AMD version 1.12 compare certaines solutions de serveur basées sur les CPU AMD EPYC™ et Intel® Xeon® requises pour offrir des PERFORMANCES TOTALES de 391000 unités de performances SPECrate2017_int_base en date du 10 octobre 2024. Cette estimation compare un serveur existant équipé de deux processeurs Intel Xeon 28 cœurs Platinum_8280 avec un score de 391 à un serveur basé sur deux processeurs EPYC 9965 (192 cœurs) avec un score de 3 000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) ainsi qu'une mise à niveau de serveur équipé de deux processeurs Intel Xeon Platinum 8592+ (64 cœurs) avec un score de 1 130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Le score réel SPECrate®2017_int_base pour le serveur 2P EPYC 9965 varie en fonction des publications des OEM.
   Estimations de l'impact environnemental effectuées en transposant ces données au moyen des facteurs électriques spécifiques par pays et région, d'après le rapport « 2024 International Country Specific Electricity Factors 10 – juillet 2024 », et l'outil de calcul des équivalences de gaz à effet de serre de l'agence de protection de l'environnement des États-Unis (United States Environmental Protection Agency Greenhouse Gas Equivalencies Calculator).
   Pour en savoir plus, rendez-vous sur https://www.amd.com/fr/legal/claims/epyc.html#q=SP9xxTCO-002A.
2. 9xx5-012 : résultats de débit de taille d'instance 32 cœurs multi-instances TPCxAI à SF30 basés sur les tests internes d'AMD en date du 05/09/2024 exécutant plusieurs instances de machine virtuelle. Le test de débit de l'IA global de bout en bout est dérivé du benchmark TPCx-AI et n'est donc pas comparable aux résultats publiés de TPCx-AI, car les résultats du test de débit de l'IA de bout en bout ne sont pas conformes à la spécification TPCx-AI.
   2P AMD EPYC 9965 (384 cœurs au total), 12 instances 32 cœurs, NPS1, 1,5 To 24x64 Go de DDR5-6400 (à 6 000 MT/s), 1 DPC, PCIe NetXtreme BCM5720 Gigabit Ethernet 1 Gbit/s, NVMe® Samsung MZWLO3T8HCLS-00A07 3,5 To, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (performance de débit du profil tuned-adm, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé)
   2P AMD EPYC 9755 (256 cœurs au total), 8 instances 32 cœurs, NPS1, 1,5 To 24x64 Go de DDR5-6400 (à 6 000 MT/s), 1 DPC, PCIe NetXtreme BCM5720 Gigabit Ethernet 1 Gbit/s, NVMe® Samsung MZWLO3T8HCLS-00A07 3,5 To, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (performance de débit du profil tuned-adm, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé)
   2P AMD EPYC 9654 (192 cœurs au total), 6 instances 32 cœurs, NPS1, 1,5 To 24x64 Go de DDR5-4800, 1 DPC, NVMe Samsung MZQL21T9HCJR-00A07 2 x 1,92 To, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=désactivé, déterminisme=puissance)
   Par rapport à 2P Xeon Platinum 8592+ (128 cœurs au total), 4 instances 32 cœurs, AMX activé, 1 To 16x64 Go de DDR5-5600, 1 DPC, PCIe NetXtreme BCM5719 Gigabit Ethernet 1 Gbit/s, NVMe KIOXIA KCMYXRUG3T84 3,84 To, Ubuntu 22.04.4 LTS 6.5.0-35-generic (performance de débit du profil tuned-adm, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé)
   Résultats :
   Médiane relative générationnelle des CPU
   Turin 192 cœurs, 12 instances 6067.531 3.775 2.278
   Turin 128 cœurs, 8 instances 4091.85 2.546 1.536
   Genoa 96 cœurs, 6 instances 2663.14 1.657 1
   EMR 64 cœurs, 4 instances 1607.417 1 NA
   Les résultats peuvent varier en fonction de facteurs tels que les configurations système, les versions software et les paramètres du BIOS. TPC, TPC Benchmark et TPC-C sont des marques commerciales de Transaction Processing Performance Council.
3. 9xx5TCO-001B : Ce scénario contient de nombreuses hypothèses et estimations et, bien que basé sur les recherches internes d'AMD et sur les meilleures approximations, il doit être considéré comme un exemple à titre informatif uniquement et non utilisé comme une base pour la prise de décision à la place de tests réels. L'outil d'estimation du TCO (coût total de possession) des émissions de gaz à effet de serre et des serveurs AMD version 1.12 compare certaines solutions de serveur basées sur les CPU AMD EPYC™ et Intel® Xeon® requises pour offrir des PERFORMANCES TOTALES de 39 100 unités de performances SPECrate2017_int_base en date du 10 octobre 2024. Ce scénario compare un serveur existant équipé de deux processeurs Intel Xeon Platinum_8280 28 cœurs avec un score de 391 à un serveur basé sur deux processeurs EPYC 9965 (192 cœurs) avec un score de 3 000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) ainsi qu'une mise à niveau de serveur équipé de deux processeurs Intel Xeon Platinum 8592+ (64 cœurs) avec un score de 1 130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Le score réel SPECrate®2017_int_base pour le serveur 2P EPYC 9965 varie en fonction des publications des OEM. Estimations de l'impact environnemental effectuées en transposant ces données au moyen des facteurs électriques spécifiques par pays et région, d'après le rapport « 2024 International Country Specific Electricity Factors 10 – juillet 2024 », et l'outil de calcul des équivalences de gaz à effet de serre de l'agence de protection de l'environnement des États-Unis (United States Environmental Protection Agency Greenhouse Gas Equivalencies Calculator).
4. 9xx5-040A : résultats de débit XGBoost (exécutions/heure) basés sur les tests internes d'AMD en date du 05/09/2024.
   Configurations XGBoost : v2.2.1, ensemble de données Higgs, instances 32 cœurs, FP32
   2P AMD EPYC 9965 (384 cœurs au total), 12 instances 32 cœurs, 1,5 To 24x64 Go DDR5-6400 (à 6 000 MT/s), PCIe NetXtreme BCM5720 Gigabit Ethernet 1,0 Gbit/s, 3,5 To Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (performance de débit du profil tuned-adm, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé), NPS=1
   2P AMD EPYC 9755 (256 cœurs au total), 1,5 To 24x64 Go DDR5-6400 (à 6 000 MT/s), 1 DPC, PCIe NetXtreme BCM5720 Gigabit Ethernet 1,0 Gbit/s, 3,5 To Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (performance de débit du profil tuned-adm, ulimit -l 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé), NPS=1
   2P AMD EPYC 9654 (192 cœurs au total), 1,5 To 24x64 Go DDR5-4800, 1 DPC, 2 x 1,92 To Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (performance de débit du profil tuned-adm, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT=désactivé, déterminisme=puissance), NPS=1
   par rapport à 2P Xeon Platinum 8592+ (128 cœurs au total), AMX activé, 1 To 16x64 Go DDR5-5600, 1 DPC, PCIe NetXtreme BCM5719 Gigabit Ethernet 1,0 Gbit/s, 3,84 To KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (performance de débit du profil tuned-adm, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=désactivé, déterminisme=puissance, Turbo Boost=activé)
   Résultats :
   CPU Exécution 1 Exécution 2 Exécution 3 Médiane Débit relatif Générationnel
   2P Turin 192C, NPS1 1565,217 1537,367 1553,957 1553,957 3 2,41
   2P Turin 128C, NPS1 1103,448 1138,34 1111,969 1111,969 2,147 1,725
   2P Genoa 96C, NPS1 662,577 644,776 640,95 644,776 1,245 1
   2P EMR 64C 517,986 421,053 553,846 517,986 1 NA
   Les résultats peuvent varier en fonction de facteurs tels que les configurations système, les versions software et les paramètres du BIOS.

AMD
together we advance_

AFFIRMATIONS : *(suite)*

5. MI325-012 : Débit global d'entraînement normalisé par GPU (jetons traités par seconde) pour la génération de texte à l'aide du modèle de chat Llama2-7b exécutant Megatron-LM v0.12 (BF16) lors de l'utilisation d'une longueur de séquence maximale de 4 096 jetons, en fonction des tests internes d'AMD effectués le 04/10/2024. La taille des lots correspond au plus grand micro-lot qui peut tenir dans la mémoire du GPU pour chaque système. AMD Instinct taille de lot 8, Nvidia taille de lot 2.
   Configurations :
   Système de développement AMD : 1P AMD Ryzen 9 7950X (16 cœurs), 1 GPU AMD instinct™ MI325X (256 Go, 1 000 W), 128 Gio de mémoire, ROCm 6.3.0 (version préliminaire), Ubuntu 22.04.2 LTS avec noyau Linux 5.15.0-72-generic, PyTorch 2.4.0.
   En comparaison avec
   Nvidia DGX H200 avec 2 processeurs Intel Xeon Platinum 8468, 1 GPU Nvidia H200 (141 Go, 700 W), 2 Tio (32 DIMM, 64 Gio/DIMM), CUDA 12.6.37-1, 560.35.03, Ubuntu 22.04.5, PyTorch 2.5.0a0+872d972e41.nv24.8.
   Valeur médiane du système MI325X 12 509,82 jetons/seconde/GPU
   Valeur médiane du système H200 11 824,09 jetons/seconde/GPU
   Les résultats peuvent varier en fonction des configurations créées par les fabricants de serveurs. Les performances peuvent varier en fonction de l'utilisation de pilotes plus récents et des optimisations.
   MI325-012
6. MI325-004 : sur la base des tests effectués le 28/09/2024 par AMD Performance Labs pour mesurer le débit généré par le texte pour le modèle Mixtral-8x7B en utilisant le type de données FP16. Le test a été effectué à l'aide d'une longueur d'entrée de 128 jetons et d'une longueur de sortie de 4 096 jetons pour les configurations suivantes de l'accélérateur GPU AMD Instinct™ MI325X et de l'accélérateur GPU Nvidia H200 SXM.
   1 MI325X à 1 000 W avec performances vLLM : 4 598 (jetons de sortie/s)
   contre
   1 H200 à 700 W avec TensorRT-LLM : 2 700,7 (jetons de sortie/s)
   Configurations :
   Plateforme de référence AMD Instinct™ MI325X :
   1 CPU AMD Ryzen™ 9 7950X, 1 GPU AMD Instinct MI325X (256 Gio, 1 000 W), Ubuntu® 22.04 et ROCm™ 6.3 version préliminaire
   en comparaison avec
   la plateforme Nvidia H200 HGX :
   Supermicro SuperServer avec 2 processeurs Intel Xeon® Platinum 8468, 8 GPU Nvidia H200 (140 Go, 700 W) [seul un GPU a été utilisé dans ce test], Ubuntu 22.04, CUDA® 12.6
   Les résultats peuvent varier en fonction des configurations créées par les fabricants de serveurs. Les performances peuvent varier en fonction de l'utilisation de pilotes plus récents et des optimisations.
7. PEN-012 : mesures réalisées par AMD Performance Labs le 27 août 2024 sur les spécifications actuelles de l'accélérateur DPU AMD Pensando™ Salina, conçu avec la technologie de processus 5 nm AMD Pensando™, dont les performances de ligne sont estimées à 400 Gbit/s.
   Les résultats fournis estimés calculés pour la DPU AMD Pensando™ Elba, conçue avec la technologie de processus 7 nm AMD Pensando, ont permis d'obtenir des performances de ligne de 200 Gbit/s.
   Les résultats réels peuvent varier selon la production de silicium.
   Performances prévisionnelles de Salina :
   Bande passante : 400 Gbit/s
   Connexions par seconde : 10 millions
   Paquets par seconde : 100 millions de paquets par seconde
   Délestages du cryptage : 400 Gbit/s
   IOPS de stockage : 4 millions
   Les résultats réels et les spécifications peuvent varier selon le silicium.
8. STXP-12 : Tests réalisés en septembre 2024 par AMD Performance Labs en comparant les systèmes suivants : un HP EliteBook X G1a (14 pouces) (40 W) avec processeur AMD Ryzen AI 9 HX PRO 375, carte graphique Radeon™ 890M, 32 Go de RAM, SSD 512 Go, VBS=ACTIVÉ, Windows 11 Pro ; un Dell Latitude 7450 avec processeur Intel Core Ultra 7 165H (vPro activé), carte graphique Intel Arc, VBS=ACTIVÉ, 16 Go de RAM, SSD NVMe 512 Go, Microsoft Windows 11 Pro dans la ou les applications suivantes (en mode Performances optimales) : Cinebench R24 nT Les résultats peuvent varier en fonction des configurations créées par les fabricants de PC portables. STXP-12.
9. STXP-32 : D'après les tests internes réalisés par AMD le 23/09/2024. Résultats d'autonomie évalués lors d'une visioconférence avec Microsoft Teams de neuf participants travaillant sur batterie. Configuration de test pour les systèmes AMD et Intel fonctionnant à partir d'un niveau de puissance à 90 % > 45 % à une luminosité de 150 nits et un mode de puissance réglé sur « meilleure efficacité énergétique ».
   Configuration système : HP EliteBook X G1a (14 pouces) avec processeur AMD Ryzen AI 9 HX PRO 375 (40 W), cœurs graphiques Radeon™ 890M, 32 Go de RAM, SSD 512 Go, VBS=ACTIVÉ, Windows 11 Pro. Configuration système : Apple MacBook Pro 14 avec processeur M3 Pro 12 cœurs, carte graphique intégrée Apple, 36 Go de RAM, SSD 1 To, MacOS 15.0. Configuration système : Dell Latitude 7450 avec processeur Intel Core Ultra 7 165H (28 W) (vPro activé), carte graphique Intel Arc, VBS=ACTIVÉ, 16 Go de RAM, SSD NVMe 512 Go, Windows 11 Pro.
   Les résultats peuvent varier en fonction des configurations utilisées par les fabricants de PC. Les performances peuvent également varier en fonction de l'utilisation des pilotes les plus récents. STXP-32.
10. 9xx5TCO-001C : Ce scénario contient de nombreuses hypothèses et estimations et, bien que basé sur les recherches internes d'AMD et sur les meilleures approximations, il doit être considéré comme un exemple à titre informatif uniquement et non utilisé comme une base pour la prise de décision à la place de tests réels. L'outil d'estimation du TCO (coût total de possession) des émissions de gaz à effet de serre et des serveurs AMD version 1.12 compare certaines solutions de serveur basées sur les CPU AMD EPYC™ et Intel® Xeon® requises pour offrir des PERFORMANCES TOTALES de 39 100 unités de performances SPECrate2017_int_base en date du 10 octobre 2024. Ce scénario compare un serveur existant équipé de deux processeurs Intel Xeon Platinum_8280 28 cœurs avec un score de 391 à un serveur basé sur deux processeurs EPYC 9965 (192 cœurs) avec un score de 3 000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) ainsi qu'une mise à niveau de serveur équipé de deux processeurs Intel Xeon Platinum 8592+ (64 cœurs) avec un score de 1 130 (https://spec.org/cpu2017/results/res2024q3/cpu2017-20240701-43948.pdf). Le score réel SPECrate®2017_int_base pour le serveur 2P EPYC 9965 varie en fonction des publications des OEM.
    Estimations de l'impact environnemental effectuées en transposant ces données au moyen des facteurs électriques spécifiques par pays et région, d'après le rapport « 2024 International Country Specific Electricity Factors 10 – juillet 2024 », et l'outil de calcul des équivalences de gaz à effet de serre de l'agence de protection de l'environnement des États-Unis (United States Environmental Protection Agency Greenhouse Gas Equivalencies Calculator).
    Pour obtenir plus de détails, rendez-vous sur https://www.amd.com/fr/legal/claims/epyc.html#q=epyc5#9xx5TCO-001B

AMD
together we advance_