DELL Technologies

AMD
together we advance_

2024 | Livre blanc

Renforcer le potentiel des entreprises avec l'IA : Bienvenue dans l'ère du choix

Sommaire

L'IA comme opportunité d'une profonde transformation de nombreux secteurs

Aujourd'hui est le meilleur moment pour transformer votre entreprise et la préparer aux innovations de demain grâce à l'IA (intelligence artificielle). D'après les données recueillies dans le cadre de l'étude de vision technologique 2023 d'Accenture, 98 % des dirigeants mondiaux s'accordent à dire que les modèles basés sur l'IA joueront un rôle prépondérant dans la stratégie de leur organisation au cours des trois à cinq années à venir.[1]

L'IA est devenue extrêmement utile pour les entreprises dans des domaines tels que la vente au détail, les services de santé et les services financiers, du fait de sa capacité à optimiser l'efficacité des tâches, stimuler l'innovation et améliorer les processus de prise de décision. Toutefois, malgré les avantages qu'elle présente, des obstacles à son intégration semblent persister du fait de certaines idées fausses.

Se mettre à l'IA demande une équipe de développeurs dédiés :

Si une expertise en science des données reste précieuse pour développer des solutions d'IA avancées et en comprendre les principes sous-jacents, elle n'est plus une condition préalable. La prolifération d'outils d'IA intuitifs, de plateformes telles que Hugging Face et de modèles spécifiques à certaines tâches élimine une grande partie de la complexité inhérente au développement de solutions d'IA.

Il faut investir des dizaines de millions en matériel pour obtenir des résultats :

Cette idée fausse nuit gravement à la diversité des ressources d'IA disponibles à l'heure actuelle. Si ces ressources bien connues sont souvent puissantes et bénéficient d'une bonne prise en charge, elles ne constituent pas toujours le choix le plus approprié ou le plus rentable pour chaque entreprise.

Tous les efforts doivent viser à l'obtention d'accélérateurs :

Si les accélérateurs excellent sur les lourdes charges applicatives d'IA, les entreprises n'ont pas forcément besoin d'une telle puissance de calcul pour leurs applications d'IA. Attendre indéfiniment d'avoir accès aux accélérateurs les plus performants du marché n'est tout simplement pas réaliste. Bien souvent, les processeurs optimisés par l'IA offrent les performances et l'efficacité nécessaires à la production d'analyses et de décisions assistées par l'IA en temps réel. Ils constituent dès lors une solution beaucoup plus rentable et adaptable.

[1] Accenture, 30 mars 2023, « Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds », https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

Fort heureusement, l'environnement de l'IA évolue. **Dell** et **AMD** s'associent pour faire tomber ces idées reçues en mettant des technologies et outils d'IA à la portée d'un plus grand nombre d'utilisateurs, grâce à une infrastructure de bout en bout conçue pour répondre aux exigences d'aujourd'hui en matière d'IA.

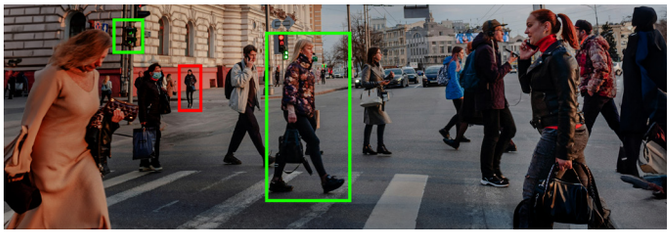
Il est possible de commencer avec un modèle déjà optimisé, une pile logicielle fiable et un système matériel polyvalent, le tout grâce au partenariat entre Dell et AMD. Plus besoin d'avoir accès à des accélérateurs de plus en plus rares, de constituer un vaste groupe d'ingénieurs en IA qualifiés ou de mobiliser d'importantes ressources en vue du déploiement de clusters Cloud massifs, pour tirer parti du potentiel de l'IA.

La collaboration entre **Dell** et **AMD** offre un écosystème matériel et logiciel unifié, conçu pour permettre aux développeurs de créer des solutions d'IA de bout en bout intégrant efficacement et en toute simplicité l'apprentissage par transfert, le réglage fin et l'inférence. Grâce au soutien de **Hugging Face**, nous disposons désormais d'une gamme en pleine expansion de modèles fonctionnant sur serveurs Dell PowerEdge équipés de processeurs AMD EPYC™ ou d'accélérateurs AMD Instinct™ MI300X, permettant aux développeurs de procéder à un réglage fin, d'appliquer l'apprentissage par transfert et de déployer pour inférence. Les investissements dans AMD ROCm™ et AMD ZenDNN™, ainsi que les partenariats avec les frameworks PyTorch, Tensorflow et ONNX Runtime, sont fondamentalement ce qui va permettre aux développeurs d'IA appliquée de faire l'expérience de la démocratisation de l'IA. Le schéma de pile ci-dessous détaille les composants de l'écosystème d'IA unifié de Dell et AMD.

La présence de l'IA selon les secteurs d'activité

Avec la diversification des ressources et l'accent mis sur l'innovation Open Source, l'IA opère actuellement une migration vers nombre de secteurs d'activité, parmi lesquels le service client, la finance et la banque, les services de santé et la vente au détail. Dans tous ces secteurs, cependant, l'IA permet collectivement aux organisations de libérer tout le potentiel de leurs propres données propriétaires et de repenser leurs workflows d'IA en s'attaquant à certaines fonctionnalités clés : analyse des données, automatisation, personnalisation et analytique prédictive. Par ailleurs, les bibliothèques AMD ROCm et ZenDNN accélèrent ces workflows d'IA pour fournir des résultats en temps quasi réel.

Voyons l'influence que l'IA exerce sur différents secteurs d'activité.

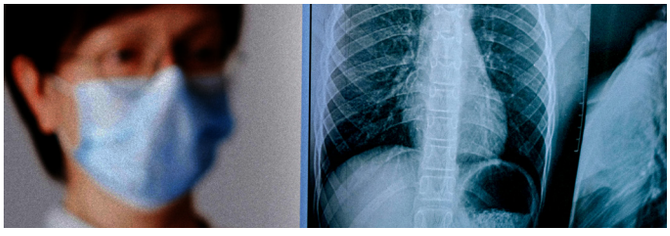


Industrie automobile

L'IA est utilisée pour la détection d'objets, le suivi des voies et la prise de décision dans les véhicules autonomes. L'IA peut aussi prédire quand une pièce du véhicule risque d'être défaillante, permettant ainsi une maintenance proactive et réduisant les interruptions de service.

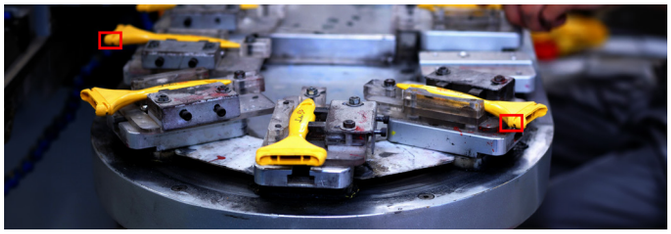
Vente au détail

L'IA peut analyser le comportement des clients en vue de recommandations de produits personnalisées, améliorant ainsi l'engagement des clients et les ventes. Elle peut aussi optimiser les niveaux de stock en prédisant la demande et en réduisant les surstocks ou les ruptures de stock.



Médical

Dans le domaine de la santé, l'IA peut couvrir diverses applications, parmi lesquelles l'analyse d'images médicales, le diagnostic de maladies, la planification de traitements personnalisés et la découverte de médicaments, permettant de meilleurs résultats pour les patients ainsi qu'une réduction des coûts.

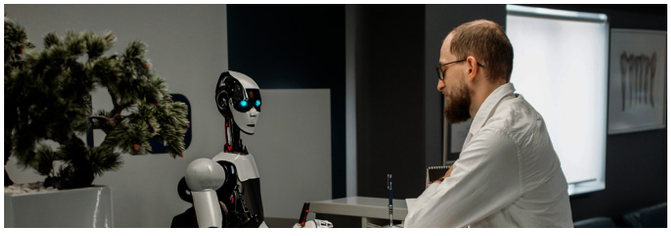


Fabrication et industrie

L'IA peut être utilisée dans la fabrication et l'industrie à des fins de maintenance prédictive, de contrôle qualité, d'optimisation des processus et de gestion de la chaîne logistique, pour une efficacité renforcée et moins d'interruptions de service.

Services financiers

L'IA peut être utilisée dans les secteurs financier et bancaire pour la détection des fraudes, l'évaluation des risques, le service client et l'analyse des investissements, renforçant la sécurité et permettant une prise de décision plus éclairée.



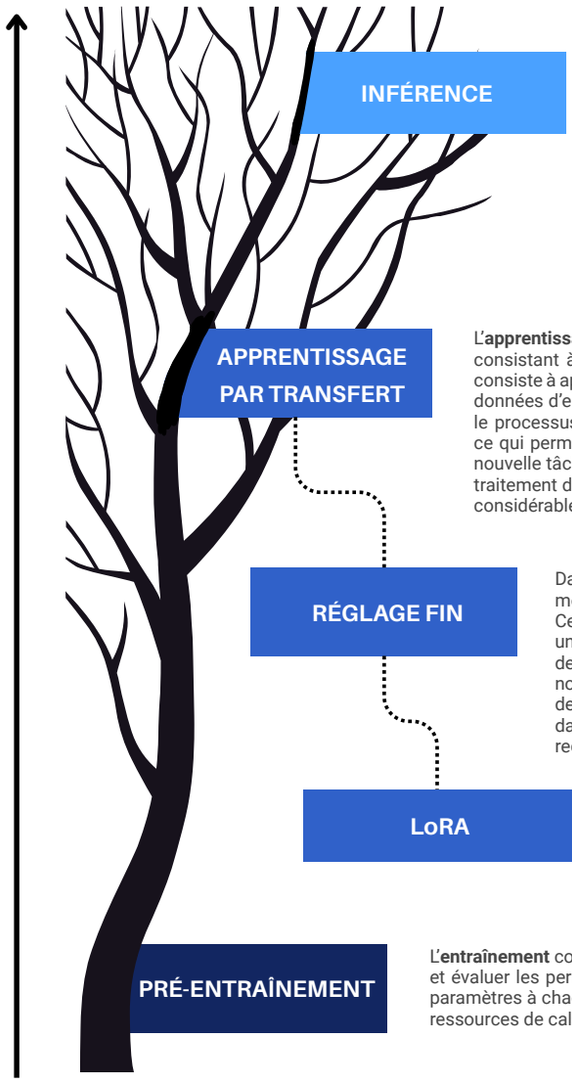
Automatisation des services

Les chatbots basés sur l'IA peuvent traiter les demandes des clients et fournir une assistance, réduisant ainsi la nécessité d'une intervention humaine. L'IA peut également automatiser les tâches répétitives telles que la saisie de données ou le traitement de documents, pour plus d'efficacité et moins d'erreurs.

Ce que les décideurs IT doivent prendre en compte

MISE EN ROUTE : DÉCOMPOSER L'IA

Avant de passer en revue ces cas d'utilisation, intéressons-nous de plus près au cycle de vie de l'IA. Le cycle de vie de l'IA désigne les étapes liées au développement, au déploiement et à la maintenance d'un système d'IA. Si les méthodologies et la terminologie spécifiques peuvent varier, un cycle de vie de l'IA type comprend toujours des phases d'entraînement et d'inférence des modèles.



L'**inférence** consiste à utiliser un modèle d'IA entraîné pour réaliser des prédictions, procéder à des classifications ou prendre des décisions sur la base de données nouvelles et inédites. À cette étape du cycle de vie de l'IA, le modèle entraîné est déployé et utilisé dans un environnement de production en vue de l'exécution de la tâche spécifique pour laquelle il a été conçu. Lors de cette phase d'inférence, le modèle prend les données d'entrée, les traite à l'aide des schémas et relations appris pendant la phase d'entraînement et génère un résultat, tel qu'une prédiction, une classification ou encore un texte, dans le cas de charges applicatives d'IA générative. Ce résultat peut ensuite être utilisé pour diverses applications telles que la recommandation de produits à un utilisateur, la détection d'anomalies dans un système ou la reconnaissance d'objets dans une image. Dans le traitement du langage naturel, la génération augmentée de récupération (RAG) améliore la précision lors du déploiement des modèles en leur permettant d'accéder à des sources de connaissances externes et de les exploiter. Les modèles comprennent alors mieux les données d'entrée en langage naturel et peuvent y répondre dans un plus grand nombre de contextes.

L'**apprentissage par transfert**, également appelé réentraînement, est une technique d'apprentissage automatique consistant à réutiliser un modèle entraîné pour une tâche en vue d'une deuxième tâche connexe. L'idée générale consiste à appliquer les connaissances qu'un modèle a acquises dans le cadre d'une tâche comportant de nombreuses données d'entraînement étiquetées, à une nouvelle tâche pour laquelle il existe peu de données. Au lieu de reprendre le processus d'apprentissage à zéro, nous partons des schémas appris lors de la résolution d'une tâche connexe, ce qui permet de mettre à jour les paramètres du modèle entraîné pour la tâche connexe pour mieux l'adapter à la nouvelle tâche. L'apprentissage par transfert est principalement utilisé dans des tâches de vision par ordinateur et de traitement du langage naturel telles que l'analyse des sentiments, afin de réduire la puissance de calcul normalement considérable qu'elles requièrent.

Dans le cadre du cycle de vie de l'IA, le **réglage fin** renvoie au processus d'entraînement supplémentaire d'un modèle d'apprentissage automatique pré-entraîné sur un nouveau jeu de données ou une nouvelle tâche. Cette technique est couramment utilisée dans l'apprentissage par transfert, lorsqu'un modèle entraîné sur un vaste jeu de données pour une tâche spécifique est adapté à une tâche connexe ou à un jeu de données de taille plus réduite. Le réglage fin permet de transférer les connaissances du modèle pré-entraîné vers la nouvelle tâche ou le nouveau jeu de données, ce qui se traduit souvent par une convergence plus rapide et des performances accrues par rapport à l'entraînement d'un modèle à partir de zéro. Il est largement utilisé dans diverses applications d'IA, notamment la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale.

Le **LoRA** (Low Rank Adaptation) est une nouvelle technique de réglage fin des modèles de Deep Learning qui consiste à réduire le nombre de paramètres entraînables et permet un changement de tâche efficace. Le processus de réglage fin consomme alors moins de ressources de calcul et est plus accessible grâce à du matériel largement disponible.

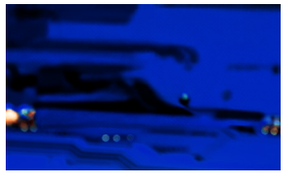
L'**entraînement** consiste à sélectionner l'algorithme ou l'architecture de modèle approprié, régler les paramètres du modèle et évaluer les performances du modèle en interne. Selon la taille du modèle en cours d'entraînement, la mise à jour des paramètres à chaque itération du processus d'entraînement peut nécessiter de gros volumes de données et d'importantes ressources de calcul.

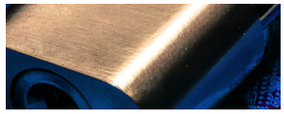
Figure 1 : Cycle de vie de l'IA

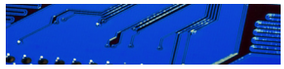
CHOIX STRATÉGIQUES

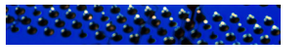
| Performances

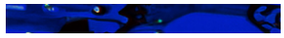
La prise de décision en temps réel ou quasi réel est essentielle à la réussite de nombre d'applications concrètes. Par exemple, il est impératif d'identifier rapidement toute activité frauduleuse dans le cadre de transactions financières ou de demandes d'indemnisation auprès des assurances afin d'éviter toute perte d'argent et de protéger les actifs des entreprises. Dans un contexte de fabrication, les défauts dans la chaîne de montage ou les conditions au sein de l'usine doivent faire l'objet d'un suivi dynamique dans un souci d'assurance qualité. En effet, le processeur traitant votre charge applicative d'inférence doit être optimisé de manière à traiter rapidement et efficacement les flux de données entrants. L'association des serveurs Dell PowerEdge aux processeurs AMD EPYC offre une polyvalence convenant parfaitement à la gestion des charges applicatives d'inférence, ainsi qu'aux tâches impliquant le calcul haute performance, le Cloud Computing et l'analytique du Big Data.











| Sécurité des données

La **sécurité des données** joue un rôle crucial dans la réussite des systèmes d'IA, en particulier ceux recourant à l'IA générative. Elle constitue une préoccupation majeure pour les leaders technologiques désireux d'intégrer l'IA dans leurs opérations. Les systèmes d'IA s'appuient généralement sur des quantités massives de données, lesquelles peuvent inclure des informations sensibles et confidentielles telles que des données personnelles, financières ou exclusives. Protéger ces données est indispensable afin d'éviter tout accès non autorisé ou vol de données, ainsi que pour garantir la précision, la fiabilité et la cohérence des modèles d'IA et de leurs prédictions.

L'**informatique confidentielle** est une technologie qui facilite le traitement des données dans une enclave sécurisée en les protégeant contre les accès non autorisés ou la manipulation par des parties non autorisées, y compris le fournisseur de Cloud et d'autres utilisateurs.[2] Les données sont isolées au cours du traitement par le biais d'un chiffrement et d'autres mesures de sécurité. AMD Infinity Guard, ensemble de fonctions de sécurité sophistiquées intégrées aux processeurs AMD EPYC, prend en charge l'informatique confidentielle via Secure Encrypted Virtualization (SEV), qui chiffre les machines virtuelles (VM) à l'aide d'une clé qui n'est connue que du processeur. Ces services visent à fournir des environnements d'exécution matériels fiables à l'aide de la technologie AMD SEV-Secure Nested Paging (SEV-SNP), qui améliore la protection des invités et contribue à la défense contre les menaces externes.

L'**apprentissage fédéré** est une autre méthode permettant de garantir la sécurité des données, consistant à entraîner un modèle central sur des appareils ou des serveurs décentralisés.[3] Plutôt que de transférer toutes les données vers un emplacement central, chaque appareil entraîne le modèle localement et seules les mises à jour du modèle sont partagées. Cette approche préserve la confidentialité tout en permettant un apprentissage collaboratif sans partager les données brutes. La plateforme d'IA fédérée de Dell Technologies permet d'exécuter des processus de calcul, des algorithmes d'IA et de ML sur des jeux de données à la périphérie du réseau au fur et à mesure de leur collecte, en partageant uniquement des modèles mathématiques, des métadonnées et des résultats de requêtes sur le réseau vers d'autres appareils en périphérie, les datacenters ou le Cloud. Cet échange améliore les résultats en permettant l'extraction en temps quasi réel d'informations exploitables à partir de vastes jeux de données distribués, sans révéler les données ni aucune propriété intellectuelle.

[2] Advanced Micro Devices, Inc. 30 août 2023, « AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings », https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc., 2021, « Data Center Solutions, Confidential Computing », Présentation de solution, https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

[3] Analytics Vidhya, décembre 2023, « Federated Learning: A Beginner's Guide », https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies, 2021, « A federated learning platform for real-time artificial intelligence », Présentation de solution, https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

METTRE À L'ÉCHELLE VOTRE SOLUTION

| Trouver l'équilibre entre coût et innovation

Trouver le juste équilibre entre coût et innovation permet de s'assurer que les solutions d'IA sont non seulement financièrement réalisables, mais aussi efficaces au sens où elles génèrent une valeur réelle pour les entreprises et les utilisateurs. Pour trouver cet équilibre, il est essentiel d'identifier un matériel qui réponde à vos besoins et s'intègre facilement dans l'infrastructure existante. Le marché moderne des matériels d'IA est en proie à une pénurie d'accélérateurs, due tant à l'augmentation de la demande en accélérateurs dans divers secteurs qu'aux contraintes de capacité de production, défis logistiques et autres pénuries de semi-conducteurs.

Toutefois, les processeurs étant déjà un composant standard dans la plupart des datacenters, l'intégration s'avère plus simple et plus rentable que l'ajout d'un matériel d'accélération entièrement neuf. Les processeurs optimisés par l'IA peuvent exploiter les logiciels et outils existants, limitant ainsi la nécessité d'acquisition de nouveaux outils ou d'efforts de formation conséquents. Les processeurs offrent également une plus grande flexibilité et une meilleure efficacité pour un large éventail de tâches au-delà de l'IA, ce qui permet une utilisation plus polyvalente des ressources au sein du datacenter. L'actualisation de votre datacenter avec des serveurs Dell PowerEdge équipés de processeurs AMD EPYC permet de satisfaire vos charges applicatives existantes, tout en vous préparant aux grandes avancées en termes d'innovation et d'efficacité annoncées par l'IA.

| Simplicité et flexibilité

La simplicité et la flexibilité de votre système d'IA sont essentielles pour élaborer des solutions d'IA efficaces, adaptables et évolutives à long terme. L'accès à une suite de frameworks et d'optimisations logiciels complétant votre matériel améliore les performances sans que vous ayez besoin de consacrer du temps et des efforts supplémentaires à une intégration multiplateforme. Ces qualités sont particulièrement importantes pour traiter les charges applicatives d'IA mixtes impliquant une combinaison de différents types de tâches d'IA, telles que l'entraînement, l'inférence et le traitement des données.

AMD et Dell Technologies s'attaquent aux charges applicatives d'IA mixtes à l'aide d'une combinaison de solutions matérielles et logicielles. Les processeurs AMD EPYC offrent une puissance de calcul haute performance grâce à des fonctionnalités telles que le multithreading simultané (SMT) et un nombre élevé de cœurs, ce qui permet un traitement parallèle efficace des charges applicatives d'IA. Ces processeurs sont optimisés pour les tâches d'IA et offrent d'excellentes performances pour les charges applicatives d'entraînement et d'inférence. Les serveurs Dell PowerEdge, équipés de processeurs AMD EPYC, constituent une plateforme évolutive et flexible pour le déploiement de charges applicatives d'IA. En outre, la suite logicielle Dell OpenManage offre des outils de gestion permettant d'optimiser l'allocation de ressources et la surveillance des performances pour les charges applicatives d'IA mixtes.

AMD propose également l'Unified Inference Frontend (UIF), qui exploite les versions améliorées de chacune des piles logicielles actuelles et s'appuie sur la bibliothèque AMD ZenDNN pour les processeurs AMD EPYC, la pile AMD ROCm Open Source pour les accélérateurs AMD Instinct, ainsi qu'une pile logicielle pour les SoC adaptatifs AMD. AMD ROCm est également conçu pour fonctionner avec une large gamme de processeurs et d'accélérateurs AMD, y compris des produits professionnels et grand public.

| Garantir l'explicabilité

L'**IA explicable** joue un rôle central pour garantir la transparence, la fiabilité et l'efficacité des applications d'intelligence artificielle. L'IA explicable permet de comprendre comment les modèles d'IA prennent des décisions, en mettant en évidence les facteurs et les processus de raisonnement sous-jacents. Cette transparence est essentielle pour gagner la confiance des parties prenantes, en particulier dans des domaines sensibles comme les services de santé, la finance et la justice pénale, où les décisions ont un impact direct sur la vie des individus.

Les systèmes d'IA **Human-in-the-loop** exploitent l'intelligence humaine pour améliorer les performances de l'IA et atténuer les biais algorithmiques. En intégrant la supervision humaine, ces systèmes peuvent gérer plus efficacement des situations complexes et ambiguës, en veillant à ce que les solutions d'IA soient conformes aux normes éthiques et sociales. En outre, l'implication humaine permet d'affiner et d'adapter en continu les modèles d'IA grâce à un feedback concret, favorisant ainsi l'amélioration itérative et la fiabilité à long terme. Ces approches sont essentielles à la création de systèmes d'IA responsables et inclusifs, servant au mieux les intérêts de la société.

Scénarios concrets

Scalers AI a collaboré avec Dell et AMD pour présenter les capacités des serveurs Dell PowerEdge équipés de processeurs AMD. Découvrez comment ces technologies sont utilisées pour l'entraînement, l'apprentissage par transfert et l'inférence dans des scénarios de vente au détail et de services de santé.

VENTE AU DÉTAIL

Scalers AI a mis au point la solution de référence pour la gestion des stocks au détail, un système conçu pour surveiller et gérer les niveaux de stock dans les rayons des magasins, grâce à la mise en œuvre d'un modèle d'IA de détection d'objets. Cette solution de référence s'appuie sur le modèle SSD_MobileNet_V2 pour identifier et reconnaître les produits dans les rayons des magasins, ce qui permet à terme d'automatiser les inventaires et de surveiller avec précision les niveaux de stock. Le modèle a fait l'objet d'un apprentissage par transfert à l'aide du jeu de données d'images SKU110K, comprenant 23 000 images tirées de Roboflow. En s'appuyant sur des algorithmes de vision par ordinateur et d'apprentissage automatique, le système peut détecter les articles en faible quantité ou en rupture de stock, et alerter le personnel du magasin pour qu'il procède à un réapprovisionnement en temps voulu.

Cette solution utilise le serveur Dell PowerEdge R7615 avec le processeur AMD EPYC 9354P 32 cœurs.

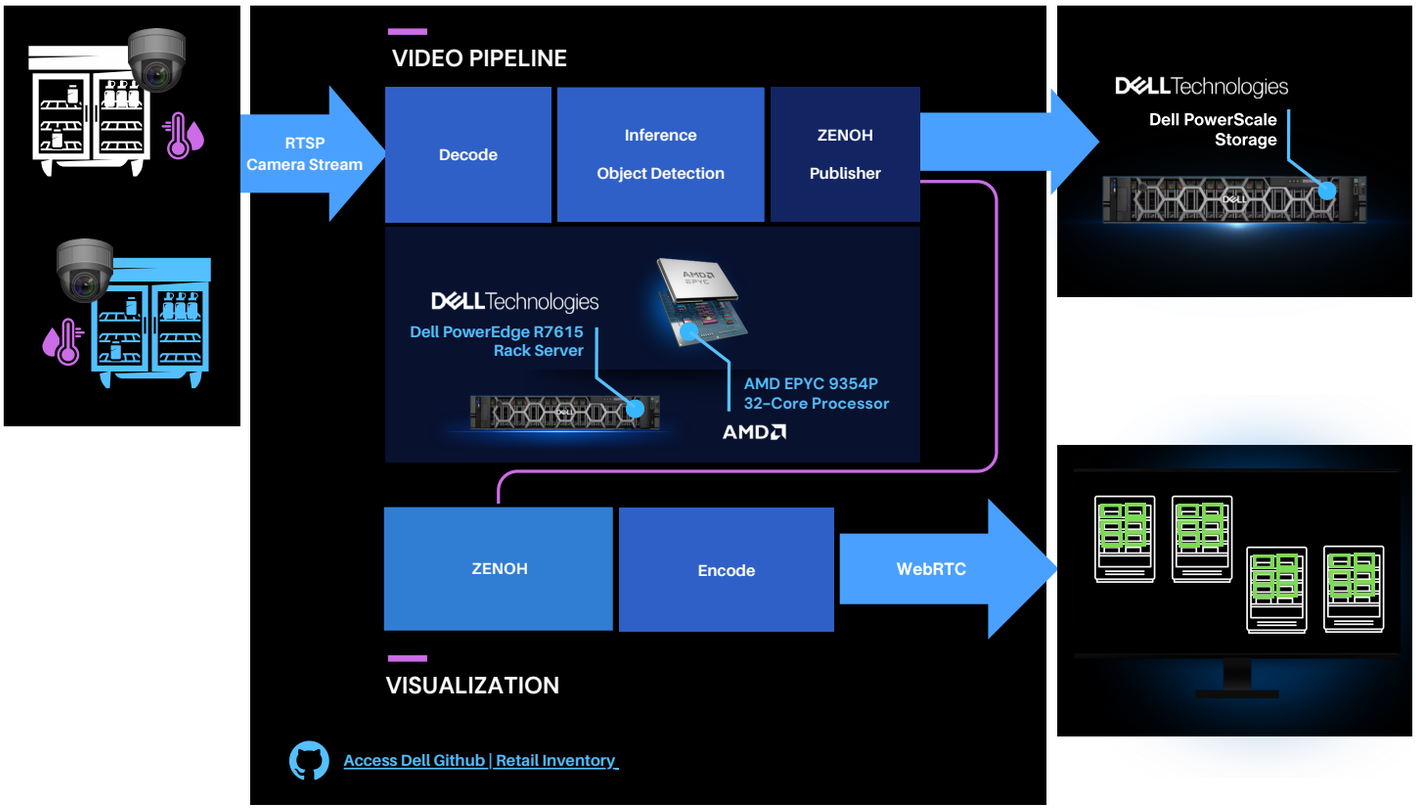


Figure 2 : Schéma de l'architecture de la solution de référence de gestion des stocks de vente au détail

SERVICES DE SANTÉ

L'imagerie médicale basée sur l'IA est extrêmement précieuse car elle peut améliorer les services de santé en augmentant la précision et l'efficacité des diagnostics, et en fournissant aux professionnels de santé des informations précises sur des affections difficilement détectables à l'œil nu. En automatisant l'analyse des images médicales, l'IA réduit le temps nécessaire au diagnostic, ce qui permet de prendre des décisions plus rapides en matière de traitement et, en définitive, d'améliorer l'état de santé des patients.

Scalers AI a exploité les capacités du serveur Dell PowerEdge R7625 équipé de processeurs AMD EPYC 9554 64 cœurs pour créer une solution d'imagerie médicale basée sur l'IA pour le dépistage de la pneumonie. En utilisant des algorithmes avancés et des techniques d'apprentissage automatique pour analyser des images médicales telles que des radiographies ou des tomodensitogrammes, la solution permet d'augmenter la vitesse et la précision du diagnostic de la pneumonie chez les patients. Une couche supplémentaire d'examen assisté par ordinateur est en fin de compte ajoutée pour aider les professionnels de santé à traiter plus efficacement d'importants volumes de données d'imagerie.

Cette solution de référence utilise le modèle ResNet50 pour analyser les images de radiographie thoracique obtenues à partir du jeu de données du NIH Clinical Center. Son objectif principal consiste à détecter la présence ou l'absence de pneumonie, en effectuant essentiellement une classification binaire. Le modèle a été entraîné à l'aide du jeu de données DICOM Xray dérivé du jeu de données du NIH Clinical Center, ce qui implique un apprentissage par transfert avec l'architecture ResNet50.

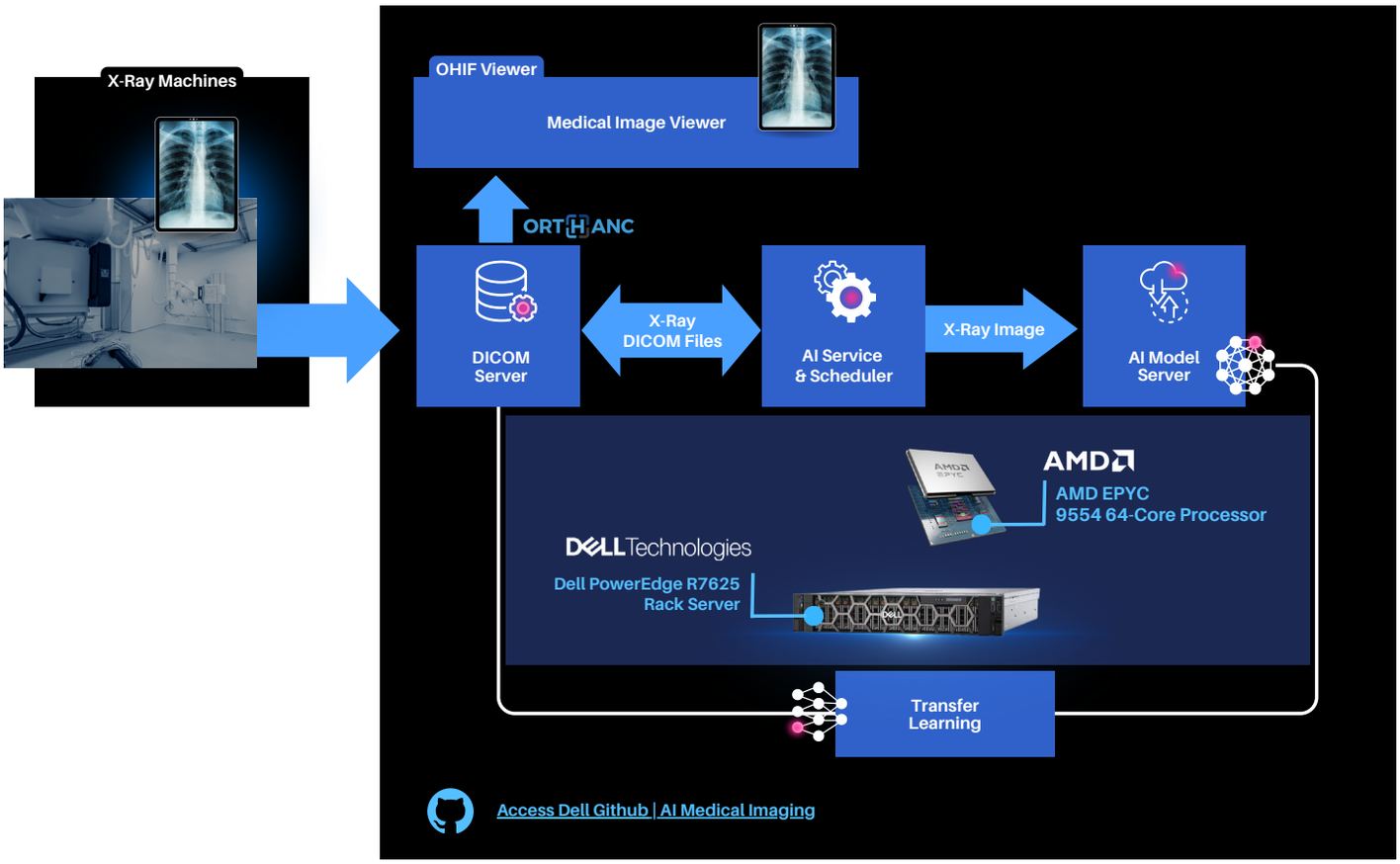
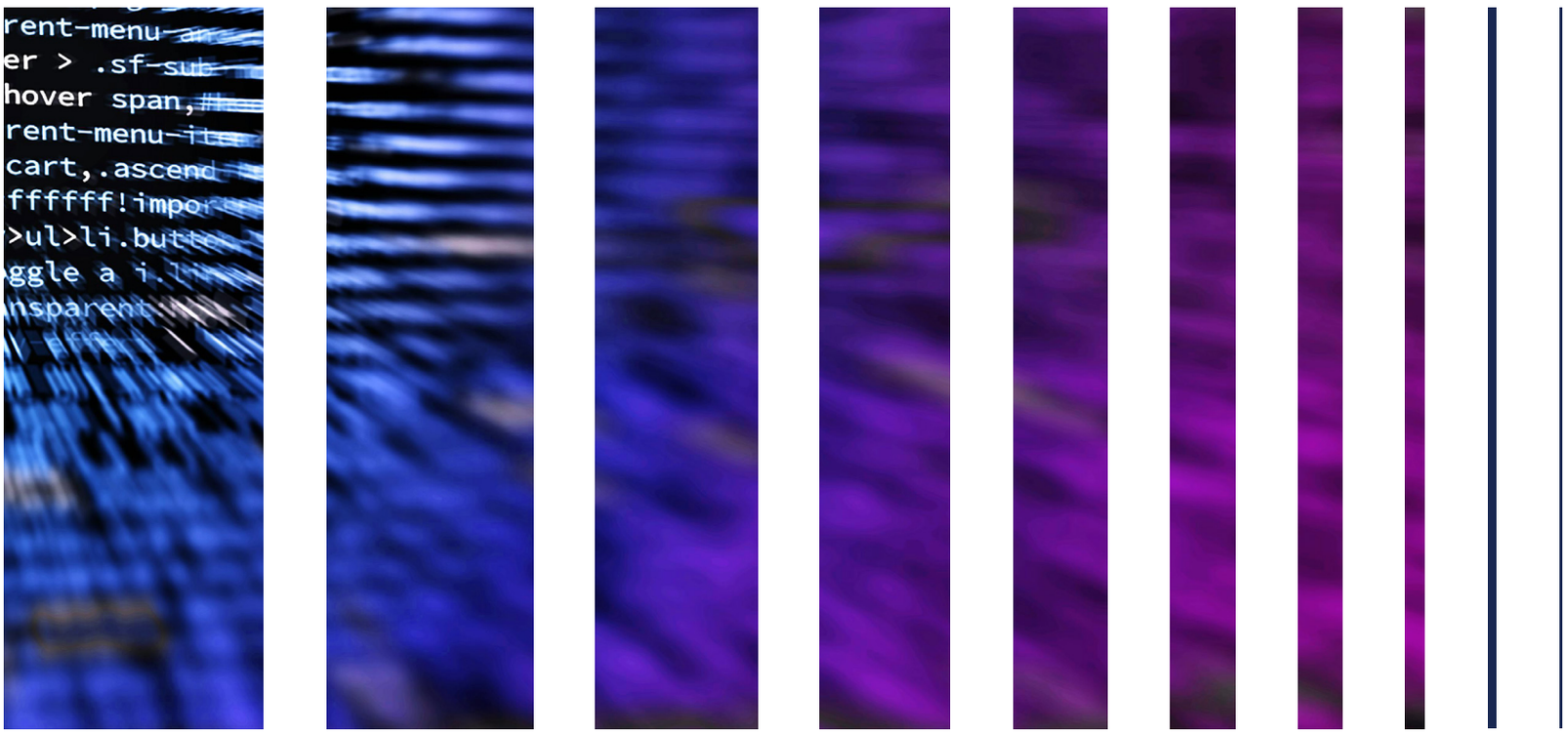


Figure 3 : Schéma de l'architecture de la solution d'imagerie médicale par IA

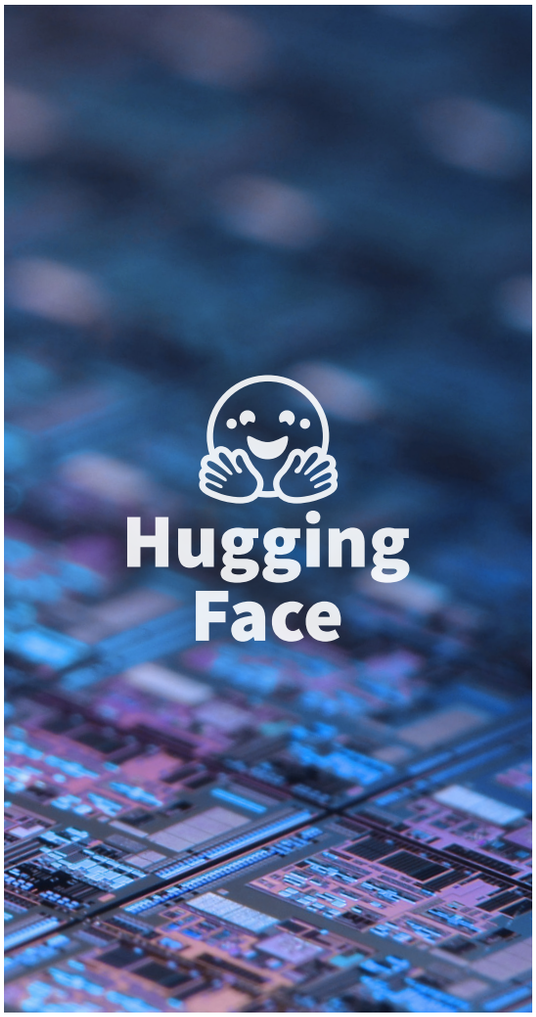
Nos solutions

L'IA S'ADRESSE À TOUS : DELL ET AMD DÉMOCRATISENT L'IA

Cette collaboration jette les bases de la démocratisation de l'IA, une étape indispensable dans la promotion de l'innovation et de l'inclusion dans l'écosystème de l'IA. Dell et AMD obtiennent ce résultat en permettant aux individus et organisations d'exploiter l'IA et de relever des défis uniques dans leurs domaines respectifs, grâce à une suite accessible de serveurs puissants équipés de technologies de processeurs et d'accélérateurs AMD de pointe. Les serveurs Dell PowerEdge équipés des accélérateurs AMD Instinct MI300X sont capables de gérer des charges applicatives d'IA volumineuses telles que l'entraînement et le réglage fin de grands modèles de langage (LLM). Les serveurs Dell PowerEdge équipés de processeurs AMD EPYC, pour leur part, excellent dans la gestion des charges applicatives d'inférence en périphérie. En plus de la plateforme matérielle sous-jacente, AMD propose la bibliothèque logicielle ZenDNN pour l'optimisation de l'inférence de Deep Learning sur les processeurs AMD, et la bibliothèque logicielle AMD ROCm pour améliorer les capacités d'entraînement, de réglage fin et d'inférence sur les accélérateurs AMD Instinct. Toutes ces options sont liées de manière transparente dans le modèle d'inférence unifié (UIF) d'AMD, grâce auquel les utilisateurs peuvent construire des solutions d'IA de bout en bout, avec une flexibilité dans le choix des frameworks, des optimisations logicielles et des plateformes matérielles.

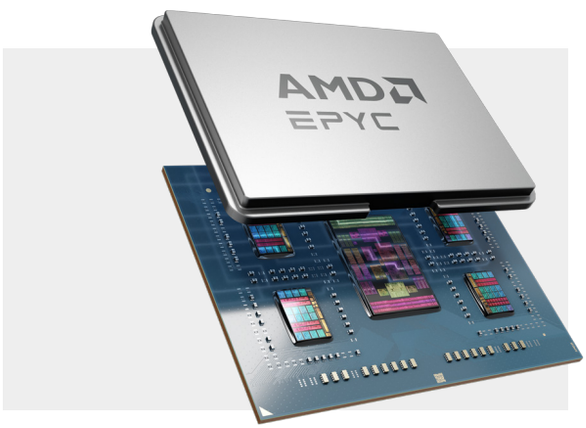


COLLABORATION AVEC HUGGING FACE



Les entreprises qui souhaitent adopter l'IA peuvent commencer par tirer parti de modèles préexistants ou de workflows d'IA adaptés à leurs besoins spécifiques directement à partir de Hugging Face, une plateforme Open Source dédiée à la science des données et à l'apprentissage automatique. AMD a conclu un partenariat avec Hugging Face dans le but commun de fournir des performances de transformateur de premier ordre, grâce à l'ajout d'optimisations logicielles spécifiques à AMD aux bibliothèques logicielles et aux frameworks qui s'intègrent déjà de manière transparente aux plateformes AMD. Hugging Face collabore activement avec l'équipe d'ingénieurs d'AMD afin d'optimiser les modèles clés pour des performances optimales, en incorporant AMD ROCm dans la bibliothèque Transformers et en améliorant Optimum-AMD, une bibliothèque spécialement conçue pour les plateformes AMD, pour aider les utilisateurs de Hugging Face à les utiliser avec un minimum de changements au niveau du code.

Dell Technologies s'est récemment associé à Hugging Face pour simplifier le processus et permettre aux entreprises de développer, d'affiner et d'appliquer leurs propres modèles d'IA générative Open Source avec l'aide de la communauté Hugging Face, le tout sur des produits et services d'infrastructure Dell leaders sur le marché. Un nouveau portail Dell est en cours de développement sur la plateforme Hugging Face. Il comprendra des conteneurs et des scripts personnalisés et dédiés qui aideront les utilisateurs à déployer sans effort et en toute sécurité des modèles Open Source disponibles sur Hugging Face à l'aide de serveurs et de systèmes de stockage de données Dell. Les entreprises peuvent désormais tirer pleinement parti des ressources de Hugging Face pour déployer directement des modèles sur des serveurs Dell PowerEdge équipés de processeurs AMD et créer des solutions d'IA de bout en bout à l'aide de leurs propres données.

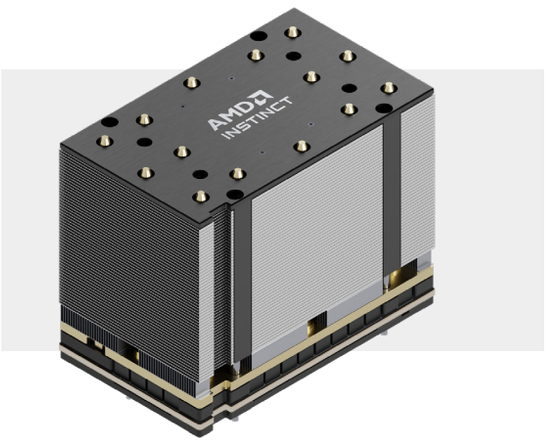


PROCESSEURS AMD EPYC

AMD offre les avancées technologiques nécessaires aux datacenters modernes basés sur le Cloud via ses processeurs AMD EPYC. Ces processeurs sont des SoC (System on Chip) conçus de A à Z pour répondre efficacement aux exigences des datacenters actuels et futurs. Les processeurs AMD EPYC série 9000 équipent le datacenter avec un maximum de 128 cœurs, 256 threads, 12 canaux de mémoire prenant en charge jusqu'à 6 To de mémoire par socket et 128 voies PCIe Gen5. Ils sont associés à la solution de sécurité pour serveurs x86 intégrée au matériel la plus innovante du marché. En intégrant des ressources essentielles de calcul, de mémoire, d'E/S et de sécurité dans le SoC, les processeurs AMD EPYC offrent des performances de premier plan et permettent de réduire le coût total de possession (TCO).

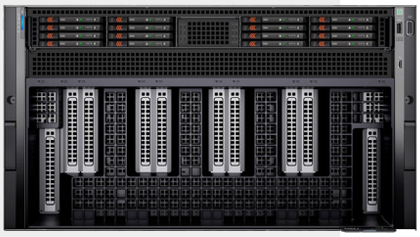
ACCÉLÉRATEURS AMD INSTINCT MI300X

L'accélérateur AMD Instinct MI300X, basé sur l'architecture de pointe AMD CDNA 3, offre une efficacité et des performances de pointe pour les applications d'IA et de HPC les plus intensives. Il est équipé de 304 unités de calcul haute performance et dispose de fonctions spécifiques à l'IA, telles que la prise en charge de nouveaux types de données et le décodage photo et vidéo, ainsi que d'une mémoire HBM3 inégalée de 192 Go sur un seul accélérateur.



PLATEFORME LOGICIELLE OPEN SOURCE AMD ROCM 6

La plateforme logicielle Open Source AMD ROCm 6 est optimisée pour maximiser les performances des accélérateurs AMD Instinct MI300X en matière de calcul haute performance (HPC) et de charges applicatives d'IA. Elle étend également la prise en charge des accélérateurs AMD Instinct MI300X, garantissant ainsi la compatibilité avec les frameworks logiciels du secteur. La plateforme AMD ROCm comprend une variété de pilotes, d'outils de développement et d'API qui facilitent la programmation des accélérateurs, du noyau jusqu'aux applications de l'utilisateur final, et peut être adaptée pour répondre à vos besoins spécifiques. La plateforme AMD ROCm est particulièrement adaptée aux applications de calcul haute performance (HPC), d'intelligence artificielle (IA) et de calcul scientifique. En outre, la plateforme AMD ROCm prend en charge le calcul multi-accélérateur, y compris l'accès direct à la mémoire à distance (RDMA) pour la communication entre le serveur et le nœud.



GAMME DE SERVEURS DELL POWEREDGE

L'investissement de Dell dans AMD introduit sur le marché un choix stratégique en vue de la démocratisation de l'IA, comme en témoignent ses quatre plateformes de serveurs avec EPYC et son serveur au format rack phare Dell PowerEdge XE9680 avec accélérateurs AMD Instinct MI300X. La dernière génération de serveurs Dell PowerEdge équipés de processeurs AMD EPYC améliore à la fois l'agilité métier et les délais de commercialisation, grâce à la prise en charge des charges applicatives de transformation, telles que les bases de données et l'analytique, la virtualisation, le stockage software-defined (SDS), l'infrastructure de bureaux virtuels (VDI), la conteneurisation, le calcul haute performance (HPC), l'IA et l'apprentissage automatique. Leurs serveurs au format rack à un socket (un seul processeur) offrent un équilibre rentable entre performances et capacité de stockage, conçus pour évoluer en toute transparence avec votre entreprise, tandis que leurs serveurs au format rack à deux sockets (deux processeurs) s'adaptent aux charges applicatives les plus exigeantes avec un large éventail de fonctionnalités.

Le serveur au format rack Dell PowerEdge XE9680 est une centrale de traitement des données robuste spécialement conçue pour les tâches d'IA. Il prend en charge huit accélérateurs, ce qui est idéal pour les charges applicatives d'entraînement et d'inférence en apprentissage automatique (ML)/Deep Learning (DL), en particulier pour ceux qui entraînent de grands modèles de langage (LLM). Équipé de huit accélérateurs MI300X, chacun doté de 192 Go de mémoire à bande passante élevée (HBM3) de 5,3 To/s, il peut atteindre une capacité HBM3 totale de 1,5 To par serveur et des performances FP16 de plus de 21 pétaflops. Le serveur au format rack Dell PowerEdge XE9680 avec accélérateurs AMD Instinct MI300X est prêt à étendre encore davantage l'accessibilité de l'IA générative aux entreprises. Cela leur permet d'entraîner des modèles plus volumineux, de minimiser l'empreinte du datacenter, de réduire le coût total de possession et d'obtenir un avantage concurrentiel.

Résumé

Le rythme effréné de l'innovation alimentée par l'IA révolutionne les charges applicatives des datacenters plus vite que toute autre transformation technologique. Pour soutenir ces avancées technologiques, Dell et AMD travaillent à la mise en place d'un écosystème d'IA plus inclusif, innovant et éthique, qui encourage les développeurs de tous les secteurs à collaborer sur des ressources Open Source et à stimuler l'innovation d'aujourd'hui en matière d'IA générative. Que votre solution d'IA réponde à vos exigences en matière de performances sur les processeurs AMD EPYC ou sur les serveurs équipés d'accélérateurs AMD Instinct, nous vous offrons la flexibilité nécessaire pour exécuter votre charge applicative d'IA sur nos plateformes matérielles, et ainsi de bénéficier du meilleur de ce que Dell et AMD ont à offrir.

CLIENT

Images AMD : AMD.com, bibliothèque de ressources partenaires AMD, **https://www.amd.com/en/partner/resources/resource-library.html**

Images Dell : **Dell.com**