

Étude de recherche technique

 PROWESS

Records mondiaux de performances pour l'IA/ML

L'étude Prowess établit une relation entre les résultats obtenus par les serveurs Dell™ à différents benchmarks et les performances réelles observées pour l'entraînement et l'inférence sur des modèles d'intelligence artificielle (IA) et d'apprentissage automatique (ML).

Synthèse

L'intelligence artificielle (IA) et l'apprentissage automatique (ML) sont utilisés dans différents secteurs d'activité pour aider les entreprises à effectuer des prévisions commerciales, à segmenter leurs clientèles, à identifier les risques, à gérer les réseaux d'approvisionnement complexes, à optimiser leurs coûts et à gagner en efficacité. Quelle que soit la manière dont ils sont employés, tous les cas d'usage IA/ML dépendent des performances de calcul (en plus de la vitesse et de la quantité de mémoire, sans oublier la bande passante des interconnexions et du trafic réseau). La sécurité est souvent tout aussi essentielle, de par la nature même des données que les entreprises utilisent pour mener certaines tâches, notamment l'analytique prédictive des clients ou la détection des fraudes.

Les technologies IA/ML nécessitent des performances de pointe, ce qui signifie que l'infrastructure (calcul, mémoire locale, bande passante réseau et stockage de données pour soutenir ces charges applicatives) peut représenter un investissement important et mérite donc une évaluation rigoureuse avant l'achat. À cet égard, les analyses comparatives (« benchmarks ») standard peuvent apporter une aide précieuse, et les records mondiaux peuvent se révéler encore plus utiles, à condition qu'ils soient évalués dans les règles de l'art.

Les entreprises doivent acquérir, stocker et traiter rapidement leurs données. Ces benchmarks peuvent apporter certains éclairages sur la rapidité de collecte, de traitement et d'accès une fois les données stockées. Afin d'évaluer la relation entre de solides performances observées au cours des benchmarks et la valeur commerciale qu'elles peuvent potentiellement produire en situation réelle, l'institut Prowess Consulting s'est penché sur la valeur que peuvent représenter d'excellents scores obtenus à des benchmarks pour les entreprises qui déploient des serveurs ayant obtenu des résultats record. En raison de sa part de marché surdimensionnée et du nombre de records mondiaux que détient Dell Technologies dans les scénarios d'IA/ML, comme le prouvent certains benchmarks reconnus dans le secteur, Prowess s'est spécifiquement intéressé aux serveurs Dell™ PowerEdge™.

Parmi les serveurs Dell PowerEdge nouvelle génération, nous en avons identifié trois qui ont récemment établi des records mondiaux au cours d'un benchmark spécialement conçu pour fournir des données de performances réelles sur les charges applicatives d'IA/ML :

- **Serveur Dell PowerEdge R6625 (record mondial, TPC Express Benchmark™ AI [TPCx-AI])**
- **Serveur Dell PowerEdge R7615 (record mondial, TPCx-AI)**
- **Serveur Dell PowerEdge R7515 (record mondial, TPC Express Benchmark™ IoT [TPCx-IoT])**

Au-delà des optimisations de performances offertes par Dell Technologies pour les charges applicatives d'IA/ML, nous avons constaté que les processeurs AMD EPYC™ et les cartes réseau Broadcom® pouvaient contribuer à optimiser les performances du Big Data et de l'analytique, à la fois pour les benchmarks et pour les applications réelles. Par exemple, la prise en charge du format numérique bfloat16 dans les processeurs AMD EPYC de 4e génération permet d'exécuter de plus grands modèles d'IA avec des datasets plus volumineux, tandis que la prise en charge du codage INT8 contribue à accélérer l'inférence sur les modèles d'IA. En outre, les doubles cartes Dell™ PowerEdge™ RAID Controller (Dell™ PERC) NVM Express® (NVMe®) et la prise en charge de PCIe® 4.0 avec les cartes réseau Broadcom permettent d'obtenir des contrôleurs d'interface réseau (NIC) 100 Gigabit Ethernet (GbE) à deux ports, éliminent les contraintes de bande passante et accélèrent encore les performances des charges applicatives d'IA/ML.

Cette étude aborde les sujets suivants :

- [**Contexte**](#)
- [**Méthodologie de recherche Prowess**](#)
- [**Benchmarks pour l'IA/ML**](#)
- [**Comment interpréter les résultats de performances**](#)



Contexte : IA et ML

L'intelligence artificielle et l'apprentissage automatique occupent désormais un rôle central dans les opérations métiers des organisations qui interviennent dans divers secteurs d'activité. Qu'il s'agisse de surveiller la fraude dans un établissement bancaire, d'établir des prévisions de ventes chez un détaillant, de chercher à obtenir des diagnostics plus précis dans le cas d'un centre hospitalier ou d'implémenter un système de maintenance prédictive sur des chaînes d'assemblage, les organisations de tous types et de toutes tailles s'appuient sur l'IA pour mettre en évidence des schémas que l'homme ne pourrait vraisemblablement pas percevoir. Les modèles d'IA en général, et les modèles de Deep Learning (DL) en particulier, sont généralement plus précis s'ils sont entraînés à partir d'un important volume de données disponibles. Cette dépendance aux données impose de déployer un environnement de stockage à la fois vaste et efficace, un système de gestion de réseau plus rapide et des serveurs plus performants, afin de produire davantage de valeur à partir de données qu'il est par ailleurs essentiel de protéger.

Les entreprises ont tendance à préférer les serveurs sur site aux implémentations Cloud, et cela pour différentes raisons. Pour les scénarios IA/ML, ce choix se justifie souvent par la gravité des données et la faible latence associées à l'utilisation de modèles d'IA. Il est souvent plus rapide et plus facile de déployer des fonctions d'entraînement d'IA au plus près des données que d'avoir à supporter le coût et le temps qu'impliquerait la migration de grandes quantités de données vers un système de calcul centralisé. Le fait d'exploiter les données là où elles résident peut également réduire la latence associée à l'entraînement d'IA, et donc accélérer l'ensemble du processus. Enfin, selon le secteur d'activité et la localisation géographique, les exigences réglementaires et les lois relatives à la souveraineté des données peuvent également suffire à convaincre une organisation de conserver ses données sur site. Dans tous les cas, les performances constituent une exigence essentielle pour les entreprises qui cherchent à traiter leurs données et à en exploiter la valeur analytique à l'aide des technologies d'IA/ML.

Les exigences de performances de certaines charges applicatives, notamment l'analytique, imposent d'adapter l'infrastructure de données afin de répondre aux contrats de niveau de service (SLA). Les interactions entre processeur, taille de la mémoire, bande passante réseau et sous-systèmes de stockage jouent un rôle essentiel. Et pour évaluer ces interactions, les résultats de benchmarks offrent un excellent moyen de comparer les performances de serveurs. Étant donné que les benchmarks produisent des résultats numériques, il peut sembler assez simple de comparer des systèmes concurrents.

Mais précisément parce qu'elles génèrent des résultats clairs et objectifs en apparence, il est absolument essentiel de comprendre ce que ces analyses cherchent à mesurer et donc ce qu'elles disent réellement des plateformes de serveurs. Les organisations qui omettent d'adopter une approche nuancée et qui privilégient aveuglément les solutions qui obtiennent les meilleurs résultats dans ces benchmarks peuvent finir par être déçues si elles constatent qu'elles n'obtiennent pas le retour sur investissement attendu.



Méthodologie de recherche Prowess

Afin d'évaluer la relation entre de solides performances observées au cours des benchmarks et la valeur commerciale qu'elles peuvent potentiellement produire en situation réelle, l'institut Prowess Consulting s'est penché sur la valeur que peuvent représenter d'excellents scores obtenus à des benchmarks pour les entreprises qui déploient des serveurs ayant obtenu des résultats record. Pour simplifier notre procédure d'enquête et nous concentrer sur la façon dont les benchmarks individuels peuvent apporter des éclairages sur les performances observées dans certains aspects de l'IA et du ML, nous nous sommes spécifiquement concentrés sur les serveurs Dell PowerEdge. Ce choix a été motivé à la fois par l'importante part de marché des serveurs Dell Technologies et par le nombre de records mondiaux décrochés par Dell Technologies dans divers benchmarks spécialement consacrés aux technologies d'IA/ML.

Un seul record mondial est déjà impressionnant en soi, mais dans le cas précis des charges applicatives d'IA/ML, il apparaît que les plateformes Dell ont atteint des records dans plusieurs benchmarks. Chacune de ces analyses peut être perçue comme une pièce du puzzle, et le cumul de plusieurs records mondiaux apporte de précieuses informations sur la façon dont les plateformes Dell peuvent fonctionner en environnement réel.

Nous avons examiné les résultats de ces benchmarks afin de déterminer quelles plateformes offraient les meilleures performances pour les différents aspects des charges applicatives d'IA/ML. Nous avons tout particulièrement concentré nos recherches sur les serveurs Dell™ montés en rack. Dell Technologies possède la plus grande part de marché de serveurs dans le monde (17,2 %) ¹ et les serveurs Dell PowerEdge sont des serveurs phares, conçus pour répondre aux besoins des charges applicatives standard à moyennes. Plus précisément, pour cette étude, nous avons examiné les plateformes en rack 1U (serveur Dell PowerEdge R6625) et 2U (serveurs Dell PowerEdge R7615 et Dell PowerEdge R7515).

Lorsqu'il s'agit d'examiner les résultats d'un benchmark, il est essentiel de tenir compte des facteurs de performances les plus importants. Pour les charges applicatives d'IA/ML, deux facteurs sont tout particulièrement essentiels :

- Performances
- Rapport prix/performances

Dell Technologies a optimisé sa plateforme PowerEdge, basée sur des processeurs AMD®, pour les solutions d'IA/ML. Les charges applicatives d'IA/ML peuvent avoir des besoins particuliers, et les serveurs PowerEdge assurent un haut degré de flexibilité qui permet aux clients d'exécuter leurs propres charges applicatives en fonction de leurs exigences spécifiques. Dell Technologies a également optimisé les performances de ses plateformes PowerEdge en intégrant des cartes PERC Dell de nouvelle génération, ainsi que des cartes réseau Broadcom NVMe qui améliorent sensiblement la bande passante.

Benchmarks pour l'IA/ML

Les benchmarks établis peuvent fournir des informations sur les utilisations courantes d'une plateforme de serveur ; ils peuvent également aider les clients à déterminer si cette plateforme répondra ou non aux besoins de ses charges applicatives. Pour cette étude, nous nous sommes spécifiquement intéressés aux benchmarks qui concernent plus directement les charges applicatives d'IA/ML.

Remarque : MLPerf™ Inference Benchmark Suite est couramment utilisé pour mesurer les performances ML. Prowess a choisi de ne pas utiliser MLPerf pour son analyse, car ce benchmark se concentre sur les performances du processeur graphique, tandis que notre étude est axée sur l'évaluation des performances économiques du processeur graphique pour l'inférence IA.

TPC Express Benchmark™ AI (TPCx-AI)	Mesure les performances de bout en bout des charges applicatives d'IA/ML représentatives du secteur
TPC Express Benchmark™ IoT (TPCx-IoT)	Mesure les performances, le rapport prix-performances et la disponibilité des systèmes qui ingèrent de grandes quantités de données à partir d'un grand nombre d'appareils

TPC Express Benchmark™ (TPCx-AI)

TPCx-AI mesure les performances d'une plateforme d'apprentissage automatique ou de science des données de bout en bout. Ce benchmark a pour but d'émuler le comportement des solutions d'IA représentatives du secteur et que l'on retrouve dans les datacenters de production et les environnements Cloud.

TPCx-AI évalue les performances de l'IA dans un certain nombre de cas d'usage. Ces cas d'usage renvoient à des problèmes uniques résolus par le pipeline de science des données DL et ML dans TPCx-AI. Le pipeline est indépendant de tout cadre ou syntaxe spécifique et peut être implémenté de différentes manières. TPCx-AI évalue généralement la génération de données, la gestion des données, l'entraînement, l'attribution de score et les phases de service.

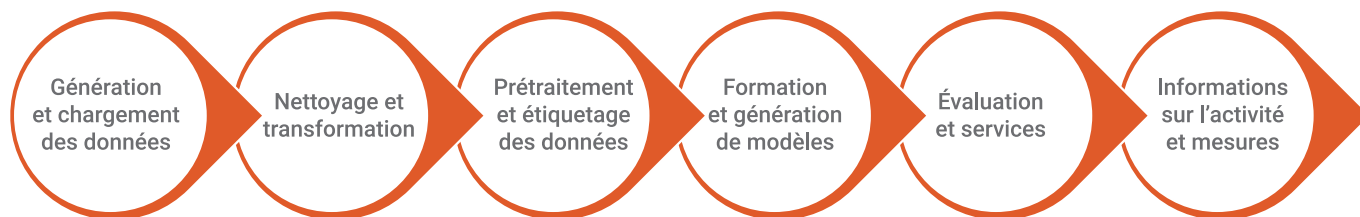


Figure 1. Le pipeline de science des données DL et ML du TPC Express Benchmark™ AI (TPCx-AI)

Ce benchmark couvre, entre autres cas d'usage, la segmentation des clients, la transcription des conversations avec les clients, les prévisions de vente, la détection des courriers électroniques indésirables, la prédiction des prix, la classification et la détection des fraudes.

Dell Technologies détient quatre records mondiaux au benchmark **TPCx-AI** avec des processeurs AMD EPYC de 4e génération :²

- Serveur Dell PowerEdge R7615 (facteur d'échelle 3)
- Serveur Dell PowerEdge R7615 (facteur d'échelle 10)
- Serveur Dell PowerEdge R6625 (facteur d'échelle 30)
- Serveur Dell PowerEdge R6625 (facteur d'échelle 100)

Les serveurs PowerEdge R7615 ont atteint un débit de pipeline IA de 408,36 Go/s au facteur d'échelle 3 (3 Go de données) et de 425,31 Go/s au facteur d'échelle 10 (10 Go de données), avec un rapport prix-performances de 118,56 Go/s au facteur d'échelle 3.²

Le serveur PowerEdge R6625 a atteint 365,59 Go/s au facteur d'échelle 30 et 868,49 Go/s dans un cluster de quatre serveurs au facteur d'échelle 100, avec des rapports prix/performances de 196,38 \$/Go/s et de 356,56 \$/Go/s, respectivement.²

Ces serveurs sont équipés de processeurs AMD EPYC 9374F, EPYC 9174F et EPYC 9354 de 4e génération. En plus d'afficher un nombre de cœurs et de threads plus élevé que la génération précédente, les processeurs AMD EPYC de 4e génération prennent en charge une mémoire DDR5 pour offrir une plus grande capacité de mémoire et le protocole PCIe Gen 5 qui assure un meilleur débit de données au niveau des serveurs et des clusters de serveurs. Ils prennent également en charge les dernières extensions vectorielles avancées 512 (AVX-512), pour un débit de données plus rapide dans le registre de processeurs, ainsi que des types numériques efficaces (notamment bfloat16 et INT8) pour accélérer l'entraînement des modèles et l'inférence. Toutes ces fonctionnalités ont aidé les serveurs Dell à chaque étape du pipeline de science des données dans le cadre du benchmark TPCx-AI.

Internet of Things (IoT) : TPC Express Benchmark™ IoT (TPCx-IoT)

L'Internet of Things (IoT) représente une grande source de données pour l'IA et le ML, en particulier pour l'inférence en temps réel par les modèles d'IA, notamment ceux utilisés pour effectuer des contrôles qualité sur les lignes de production. L'IoT représente également une source essentielle de données d'entraînement pour les modèles d'IA. Cela concerne aussi bien l'entraînement initial que les nouveaux modèles, sans oublier le réentraînement des modèles existants qui présentent des dérives.

Bien qu'il soit exclu du pipeline de science des données encapsulé dans le benchmark TPCx-AI, l'entraînement continu des modèles d'IA en production constitue un aspect essentiel. Au fil du temps, la composition des données dans les environnements de production (notamment la vente au détail, la finance ou la santé) peut évoluer de manière subtile, au point d'affecter potentiellement la précision des modèles d'IA. N'importe quel workflow d'IA/ML doit impérativement assurer une capture efficace des données d'entraînement mises à jour afin de corriger la dérive des modèles d'IA.

Ce qui différencie l'IoT des autres types de données non structurées, c'est que les appareils IoT génèrent souvent leurs données à partir de leur propre environnement, par exemple des haut-parleurs intelligents qui écoutent des commandes ou des drones industriels qui collectent des données relatives à des terres agricoles. Les données IoT doivent donc être capturées et utilisées efficacement pour les charges applicatives d'IA/ML en raison de leur volume considérable, qui est généralement supérieur à celui d'autres sources de données comme les transactions de vente au détail ou les dossiers médicaux.

Le benchmark TPCx-IoT permet de comparer directement les différentes solutions logicielles et matérielles pour les passerelles IoT. Étant donné qu'ils sont positionnés entre l'architecture de périphérie et le datacenter back-end, les systèmes de passerelle exécutent des fonctions telles que l'agrégation de données pour l'IA/ML en temps réel. Le benchmark TPCx-IoT a été spécialement conçu pour fournir des métriques de performances, de prix et de disponibilité vérifiables pour les systèmes disponibles sur le marché qui ingèrent généralement de grandes quantités de données à partir d'un grand nombre d'appareils.

Dell Technologies détient un record mondial au benchmark **TPCx-IoT** pour les serveurs Dell PowerEdge 7515 équipés de processeurs AMD EPYC de 3e génération. Le cluster a atteint un débit de 1 617 545 000 enregistrements/seconde à un rapport prix/performance de 329,75 \$/million d'enregistrements/seconde.³

Outre la puissance de traitement fournie par le processeur AMD EPYC 75F3 de 3e génération, notre analyse des résultats du benchmark indique également que les performances des serveurs Dell ont été améliorées par les cartes réseau Broadcom 25 GbE utilisées dans le cluster. Toutes ces fonctionnalités ont aidé les serveurs Dell à atteindre de solides performances au benchmark TPCx-IoT ; les résultats obtenus par Dell Technologies sont particulièrement pertinents pour les organisations qui utilisent Cloudera®, comme l'a fait Dell Technologies elle-même, pour stocker de grandes quantités de données à la périphérie pour l'analytique et l'IA.

Comment interpréter les résultats de performances

Étant donné que Dell Technologies s'est appuyée sur les processeurs AMD pour optimiser sa plateforme PowerEdge, elle a décroché de nombreux records mondiaux dans des benchmarks dédiés à l'évaluation des performances IA/ML. Dans une implémentation sur site, les processeurs AMD EPYC de 3e génération offraient déjà d'excellentes performances, de solides performances par watt et un très bon rapport prix/performance par processeur. Dans le Cloud, les SoC AMD EPYC alimentent les instances IaaS (Infrastructure as-a-service) optimisées pour le calcul haute performance (HPC) pour de nombreux prestataires de services Cloud (CSP), notamment Amazon Web Services® (AWS®), Microsoft® Azure®, Google Cloud Platform™, etc.

Les processeurs AMD EPYC de 4e génération utilisés dans les configurations de serveur record de Dell pour le TPCx-AI offrent des gains de performances qui peuvent être liés à plusieurs améliorations de la plateforme par rapport à la plateforme de génération précédente, notamment :

- 50 % d'augmentation du nombre de cœurs⁴, augmentation du nombre de threads et fréquences plus élevées, ce qui peut directement influencer sur les performances de traitement ;
- 12 barrettes DIMM/socket (jusqu'à 8), ce qui permet d'augmenter considérablement la mémoire disponible et d'accélérer ainsi le traitement de jeux de données plus volumineux, en particulier pour l'analytique en mémoire, comme ceux traités par Apache Spark™ ;
- prise en charge de la mémoire DDR5 pour un accès plus rapide aux données ;
- prise en charge d'AVX-512, qui permet aux processeurs AMD EPYC de 4e génération d'effectuer des calculs plus simultanés dans leurs registres ;
- augmentation du cache L2, qui passe de 512 Kio à 1 Mio par cœur, ce qui accélère également les opérations en mémoire ;
- prise en charge de PCIe Gen 5, qui permet des interconnexions plus rapides pour déplacer davantage de données avec une moindre latence ;
- améliorations spécifiques pour les charges applicatives d'IA et de ML, notamment prise en charge du type numérique bfloat16 afin d'accélérer l'entraînement des modèles d'IA et prise en charge de l'inférence INT8 pour augmenter les performances en production des modèles déjà entraînés.

Globalement, les processeurs AMD EPYC de 4e génération affichent une meilleure efficacité opérationnelle que leurs prédécesseurs. Les résultats du test SPEC CPU® 2017 Floating Point Rate de la Standard Performance Evaluation Corporation révèlent un gain de performances de 121 % au cours des tests réalisés sur un système équipé de processeurs AMD EPYC de 4e génération, par rapport à un système équipé de processeurs AMD EPYC de 3e génération.⁵ Les résultats du test SPEC CPU 2017 Integer Rates ont montré une amélioration de 102 %.⁶ Ces résultats de performances de processeur se reflètent à travers les records mondiaux obtenus au cours de benchmarks par plusieurs plateformes PowerEdge que nous avons examinées.

Le nombre de cœurs dans ces processeurs a augmenté de 50 % par rapport à la génération précédente, ce qui améliore également les performances. Dans le même temps, les spécifications publiées par AMD indiquent une augmentation de la consommation électrique par défaut maximale de seulement 42 %, passant de 280 watts d'enveloppe thermique (TDP) à une enveloppe thermique maximale de 400 watts.⁷ Comparativement aux résultats de performances SPEC mentionnés ci-dessus, ces chiffres de puissance thermique montrent que les serveurs équipés de processeurs AMD EPYC de 4e génération améliorent les performances d'alimentation à hauteur de 55 % pour les entreprises qui exécutent des charges applicatives d'IA/ML.⁸

Le benchmark TPCx-AI, qui mesure les performances d'IA/ML, reflète également les performances introduites par les cartes réseau Broadcom et les cartes PERC Dell. Les organisations qui déploient des charges applicatives d'IA/ML ont besoin de contrôleurs RAID pour prendre en charge la redondance afin de répondre à leurs exigences internes ou réglementaires. Les serveurs PowerEdge R6625 et PowerEdge R7615 qui ont établi les records mondiaux du TPCx-AI étaient équipés de cartes PERC Dell avec une prise en charge RAID NVMe rapide.

Sécurité basée sur le matériel AMD®

La sécurité est essentielle pour toutes les charges applicatives évaluées dans cette étude. Les processeurs AMD EPYC™ de 3e génération et AMD EPYC de 4e génération peuvent fournir une sécurité matérielle pour les charges applicatives d'IA/ML. La technologie AMD® Secure Memory Encryption (AMD® SME) chiffre la mémoire système pour protéger les données en cours d'utilisation. AMD® Secure Encrypted Virtualization (AMD® SEV) protège les machines virtuelles en cours d'exécution afin de les chiffrer et de les isoler à la fois les unes des autres et de l'hyperviseur du système hôte. AMD® Secure Encrypted Virtualization-Encrypted State (AMD® SEV-ES) chiffre le contenu du registre de processeur des machines virtuelles arrêtées afin de protéger les données qui y sont stockées. AMD® Secure Boot permet de protéger les serveurs pendant le processus de démarrage, en fournissant des moyens de défense contre les rootkits, les bootkits et le firmware, là où les serveurs sont les plus vulnérables.

Les cartes réseau de Broadcom accélèrent le flux de données pour les charges applicatives d'IA/ML. La prise en charge de PCIe Gen 5 dans les processeurs AMD EPYC de 4e génération et les cartes réseau Broadcom permet d'utiliser des cartes réseau Broadcom 100 GbE basées sur le format NIC 3.0 de l'Open Compute Project (OCP). Ces conceptions modernes reflètent une évolution rapide du secteur vers des adaptateurs 100 GbE reposant sur un format plus efficace et optimisés par les protocoles PCIe 4.0 et PCIe 5.0. En outre, la prise en charge de PCIe 4.0 et PCIe 5.0 peut, avec une seule carte réseau, délivrer des performances comparables à celles obtenues avec deux cartes NIC de 100 Gbit/s. La spécification NIC 3.0 de l'OCP permet aux fabricants de serveurs comme Dell Technologies d'utiliser des conceptions plus compactes qui peuvent prendre en charge des adaptateurs hautes performances avec des fonctionnalités avancées d'accélération matérielle pour accélérer davantage les charges applicatives d'IA/ML.⁹

Le système Dell™ PowerEdge™ RAID Controller (Dell™ PERC) protège les données et optimise les performances de stockage

Les interfaces RAID PCIe® Gen 4 modernes fonctionnent avec des disques SSD NVMe Express® (NVMe®) à large bande passante pour optimiser considérablement les performances de stockage. Les deux cartes Dell PowerEdge RAID Controller 11 et 12 (PERC 11 et PERC 12) et les adaptateurs NVMe avec des interfaces hôte PCIe Gen 4 et de stockage PCIe Gen 4 peuvent aider à éliminer les contraintes de bande passante et de latence.



Conclusion

Les résultats des benchmarks en général (et en particulier ceux qui reflètent des records mondiaux) ne se résument pas à chanter les louanges des fabricants de serveurs. S'ils sont interprétés correctement, les meilleurs résultats obtenus dans les benchmarks peuvent apporter des éclairages sur les performances des serveurs en condition d'utilisation réelle. Du fait des parts de marché et du nombre de records mondiaux affichés par Dell Technologies, les serveurs PowerEdge offrent une occasion naturelle d'examiner la corrélation entre les résultats de benchmarks et les avantages de performances pour les organisations en environnement de production. Et bien qu'il ne soit pas possible d'établir une correspondance à 1:1 pour les performances obtenues au cours de benchmarks (records mondiaux ou autres), notre enquête montre que les performances des serveurs Dell PowerEdge dans un benchmark reconnu assurent de solides performances pour plusieurs cas d'usage de l'IA/ML dans divers secteurs d'activité. En outre, le nombre de records mondiaux détenus par Dell Technologies sur différents benchmarks montre que l'entreprise a développé des plateformes capables d'exploiter les atouts de chaque composant de manière à apporter aux clients une réelle valeur ajoutée pour une grande variété de charges applicatives.

Annexe A : liens vers les performances des benchmarks

- Meilleures performances au benchmark TPCx-AI : www.tpc.org/tpcx-ai/results/tpcxai_perf_results5.asp
- Meilleures performances au benchmark TPCx-IoT V2 : www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2

Annexe B : liens vers les spécifications système Dell Technologies

- Notices techniques des serveurs Dell PowerEdge : www.dell.com/en-us/dt/servers/poweredge-rack-servers.htm

¹ History-Computer. « The 10 Largest Server Companies In The World, And What They Do. » Septembre 2022. <https://history-computer.com/largest-server-companies-in-the-world-and-what-they-do/>.

² TPC. « TPCx-AI Top Performance Results. » Consulté le 1er novembre 2022. www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2.

³ TPC. « TPCx-IoT V2 Top Performance Results. » Consulté le 1er novembre 2022. www.tpc.org/tpcx-iot/results/tpcxiot_perf_results5.asp?version=2.

⁴ Tom's Hardware. « Zen 4 Madness: AMD EPYC Genoa With 96 Cores, 12-Channel DDR5 Memory, and AVX-512. » Août 2021. www.tomshardware.com/news/zen4-madness-amd-epyc-genoa-with-96-cores-12-channel-ddr5-memory-and-avx-512.

⁵ Performances des meilleurs processeurs AMD EPYC™ de 4e génération au benchmark SPEC® Floating Point jusqu'à 121 % supérieures à celles des meilleurs processeurs AMD EPYC de 3e génération, sur la base d'un score SPEC Floating Point de 1 410 atteint sur un serveur Dell™ PowerEdge™ R7625 équipé de processeurs AMD EPYC 9654, contre un score de 636 obtenu sur un serveur Dell PowerEdge R7525 équipé de processeurs AMD EPYC 7763. Scores consultés à partir du 10 novembre 2022. Reportez-vous aux résultats du benchmark de la Standard Performance Evaluation Corporation. <http://spec.org/benchmarks.html>.

⁶ Performances des meilleurs processeurs AMD EPYC™ de 4e génération au benchmark SPEC® Integer Rate jusqu'à 102 % supérieures à celles des meilleurs processeurs AMD EPYC de 3e génération, sur la base d'un score SPEC Integer Rate de 1 660 atteint sur un serveur Dell™ PowerEdge™ R7625 équipé de processeurs AMD EPYC 9654, contre un score de 821 obtenu sur un serveur Dell PowerEdge R7525 équipé de processeurs AMD EPYC 7763. Scores consultés à partir du 10 novembre 2022. Reportez-vous aux résultats du benchmark de la Standard Performance Evaluation Corporation. <http://spec.org/benchmarks.html>.

⁷ AMD. Page Web des caractéristiques des processeurs AMD EPYC série 7003. www.amd.com/en/processors/epyc-7003-series.

⁸ 55 % d'amélioration des performances processeur par watt, calculée à partir du score SPEC® Floating Point de 1 410 atteint sur un serveur Dell™ PowerEdge™ R7625 équipé de processeurs AMD EPYC 9654 avec un cTDP de 400 watts, contre un score de 636 obtenu sur un serveur Dell PowerEdge R7525 équipé de processeurs AMD EPYC 7763 avec un cTDP de 280 watts.

⁹ Broadcom. Fiche produit des adaptateurs Ethernet NetXtreme E-Series OCP NIC 3.0. 2021. <https://docs.broadcom.com/doc/12395120>.

