RAPPORT D'ANALYSTE

Sponsorisé par : Dell

Les investissements dans l'infrastructure de traitement de l'IA augmentent à un rythme accéléré. La bonne nouvelle est qu'en ce qui concerne l'infrastructure de traitement de l'IA, plus de 50 % des systèmes ne seront pas accélérés en 2024 et pourront s'exécuter sur des serveurs standard et sur un réseau Ethernet.

L'IA générative évolue à une vitesse record alors que les entreprises entament leur transition vers l'IA générative

Février 2024

Écrit par : Brandon Hoff, directeur de recherche, Technologies habilitantes : gestion de réseau et communication, et Vijay Bhagavath, vice-président de la recherche, Réseaux Cloud et datacenter

Introduction

L'étude IDC fournit des prévisions et des facteurs sous-jacents qui devraient avoir un impact sur les investissements informatiques en 2024 et au-delà. Les leaders technologiques et leurs homologues des lignes de produits (LOB) peuvent utiliser ce document pour orienter leurs efforts de planification stratégique.

Les équipes opérationnelles capturent des données, créent des Data Lakes et exploitent le Cloud pour stocker leurs données. Aujourd'hui, avec la popularité de ChatGPT, le moment iPhone pour l'IA générative (GenAI), les équipes opérationnelles savent ce qu'elles peuvent faire avec leurs ensembles de données. Comme tout le monde connaît les avantages de l'IA générative, les équipes opérationnelles sont également confrontées à une pression supplémentaire de la part des investisseurs, des dirigeants et du marché pour mettre en œuvre une stratégie GenAI efficace. Il existe plusieurs technologies et un large éventail d'options qui peuvent être utilisées pour améliorer les opérations commerciales et la productivité des collaborateurs, de l'IA générative au ML, en passant par les jumeaux numériques et bien plus encore. La mise en œuvre réussie de la bonne technologie deviendra un KPI essentiel pour les équipes opérationnelles et l'entreprise dans son ensemble.

EN BREF

ÉLÉMENTS CLÉS

Commencez à planifier l'infrastructure GenAI :

- Intégrez l'IA plus rapidement que les autres, tout d'abord, en déplaçant au moins une copie des données sur site via une initiative visant à extraire les données du Cloud.
- Investissez dans la compréhension de la valeur des algorithmes qui sous-tendent l'IA générative, l'IA, le ML et les jumeaux numériques dans l'entreprise, et hiérarchisez en fonction de la valeur commerciale.
- Trois étapes : déploiement de serveurs standard et de la gestion de réseau Ethernet pour l'évaluation de GenAI. Faites évoluer l'IA générative pour les charges applicatives de taille entreprise en fonction des besoins, grâce à l'Ethernet standard. Rééquilibrer les charges applicatives entre l'infrastructure sur site et hors site afin d'optimiser les CAPEX et les OPEX au cours des trois à cinq prochaines années.

Comprendre l'IA générative aujourd'hui

Face à l'explosion de la demande d'IA, les prestataires de services Cloud (SP) et les entreprises développent leur infrastructure à un rythme accéléré. Les prestataires de services Cloud consomment la grande majorité des accélérateurs d'IA et créent leur propre infrastructure d'IA, mais ces accélérateurs sont coûteux, ce qui fait grimper le coût des services d'IA dans les prestataires de services Cloud. Les accélérateurs d'IA se composent de processeurs graphiques, de TPU, de FPGA, d'ASSP et d'ASIC. Ces prestataires de services Cloud construisent des usines d'IA pour des charges applicatives massives qui couvrent les besoins d'un large éventail d'entreprises et sont principalement déployées dans les plus grands environnements informatiques du monde, soit environ neuf entreprises.

D'autre part, l'infrastructure GenAI pour les charges applicatives de la taille de l'entreprise peut être construite avec des systèmes standard sans accélération requise. IDC prévoit que plus de 50 % des systèmes d'IA générative ne seront pas accélérés en 2024. Par conséquent, n'importe qui peut commencer à déployer son infrastructure GenAI avec des serveurs et une gestion de réseau standard. Des processeurs graphiques sont également disponibles, pour ceux qui en ont besoin. Il existe plusieurs options pour déployer une infrastructure d'IA, ainsi que plusieurs types d'IA générative, d'IA, d'apprentissage automatique et de jumeaux numériques qui profiteront à différentes entreprises de différentes manières. Il y a des avantages à exécuter GenAI sur des serveurs standard, car les piles logicielles GenAI sont généralement prises en charge. Les entreprises qui investissent dans leur infrastructure standard sur site seront en mesure de faire progresser leurs initiatives d'IA générative plus rapidement que les autres. Il est essentiel que les équipes informatiques commencent à évaluer les différents algorithmes d'IA générative, d'IA, d'apprentissage automatique et de jumeau numérique afin d'identifier ceux qui ont le plus d'impact sur leur entreprise.

Avantages

L'IA générative et d'autres modèles fondamentaux changent la donne en portant les technologies d'assistance à un niveau inédit et en apportant de puissantes fonctionnalités aux utilisateurs non techniques. L'IA générative a le potentiel d'accroître l'efficacité et la productivité, d'ouvrir de nouvelles opportunités de croissance, de réduire les coûts et de fournir un avantage concurrentiel aux entreprises qui en tirent parti.

La création de votre propre infrastructure d'IA générative permet d'intégrer cette technologie révolution-naire dans les opérations métier et d'acquérir une expertise sur site dans la pile de technologies d'IA générative. En donnant la priorité aux investissements technologiques qui construisent l'infrastructure initiale d'IA générative sur site, sur la base de serveurs d'entreprise standard et d'un réseau Ethernet, les entreprises qui tirent parti de cette technologie transformatrice bénéficieront d'un délai de commercialisation.

Points à prendre en compte

Agissez dès maintenant pour tirer parti de l'IA générative

Il y a de l'excitation autour de la GenAI, étant donné les résultats impressionnants que ChatGPT et d'autres modèles fournissent, GenAI apporte de la valeur, mais la valeur variera en fonction de la source des données propriétaires et des algorithmes déployés. Les conseils d'administration, les investisseurs et les dirigeants poseront des questions et chercheront à voir comment l'IA générative peut aider leur entreprise.

Pour les entreprises qui ont capturé de grandes quantités de données propriétaires non structurées, GenAI promet de créer du contenu original à partir des données propriétaires existantes qui devrait aider à recâbler l'organisation pour une innovation continue. Il est judicieux d'adopter une approche « rampant, marchant » et « courant » pour comprendre ce que l'IA générative peut apporter à l'entreprise et comment aller de l'avant.

Création de votre infrastructure initiale d'IA générative sur Ethernet

Pour les charges applicatives de la taille de l'entreprise, les systèmes standard fournissent les performances nécessaires pour entamer la transition vers l'IA générative. De plus, le fait de baser l'infrastructure GenAI sur des serveurs standard et la gestion de réseau Ethernet permet d'utiliser des systèmes d'exploitation d'entreprise, des outils de gestion d'entreprise et des outils de gestion de réseau d'entreprise. Une fois que les exigences de calcul pour les LLM qui profitent à l'entreprise sont comprises, les performances de calcul peuvent être améliorées avec la bonne sélection d'accélérateurs d'IA générative et d'IA. L'essentiel est que l'infrastructure d'IA soit bien conçue. Un fabric bien conçu peut prendre en charge des dizaines à des milliers de nœuds de calcul d'IA.

Bien qu'il puisse exister différentes options de gestion de réseau pour les charges applicatives d'IA générative, la gestion de réseau Ethernet est l'option omniprésente, ouverte et multifournisseur. Les déploiements initiaux de GenAI peuvent être pris en charge avec la mise en réseau Ethernet standard disponible dès aujourd'hui pour les clusters GenAI.

Création d'une infrastructure d'IA générative avec Ultra Ethernet

À mesure que chaque entreprise entame sa phase de développement de l'IA générative, il peut être judicieux de créer sa propre infrastructure d'IA générative scale-out. La création d'une infrastructure GenAI scale-out nécessite deux ajouts clés : des accélérateurs d'IA de datacenter et une mise en réseau de l'IA. Dans une infrastructure GenAI scale-out classique, huit processeurs graphiques de datacenter sont déployés sur chaque serveur, et pour chaque processeur graphique, une carte NIC ou un DPU haut débit est déployé pour fournir une gestion de réseau hautes performances.

« Le marché de la commutation Ethernet de centre de données GenAI dans le segment des entreprises devrait croître à un TCAC de 158,2 %, passant de 41,9 millions de dollars en 2023 à 1,0 milliard de dollars en 2027 », Vijay Bhagavath, IDC.

L'une des principales exigences d'une infrastructure d'IA scale-out est la mise en réseau hautes performances. Pour les LLM d'IA générative, le goulot d'étranglement dans le traitement est le temps que les données passent dans le réseau. Pour certaines charges applicatives, le temps passé sur le réseau peut représenter jusqu'à 60 % du temps de traitement d'un LLM, ce qui laisse l'infrastructure de calcul inactive lorsque les données se déplacent entre les clusters de calcul. Pour le réseau d'IA, il existe une mise en réseau améliorée qui est disponible aujourd'hui et fournie par le consortium Ultra Ethernet qui promet une interconnexion aussi performante que les réseaux de supercalculateurs, évolutive vers le centre de données cloud et aussi rentable et omniprésente qu'Ethernet. La gestion de réseau de l'IA est essentielle pour répondre à la croissance des demandes réseau de l'IA générative et du HPC à grande échelle. La bonne nouvelle est que le consortium Ultra Ethernet est pris en charge par la plupart des fournisseurs de commutateurs Ethernet.

Pour des performances optimales, trois technologies clés sont requises : les SerDes haut débit, les couches PHY et les optiques. Ces trois technologies sont utilisées dans Ethernet et dans d'autres technologies de gestion de réseau. Par conséquent, il n'y a fondamentalement aucun avantage en termes de performances pour une technologie de gestion de réseau spécifique. Pour obtenir les meilleures performances d'Ethernet, l'InfiniBand Trade Association a lancé l'initiative RDMA over Converged Ethernet (RoCE) et défini le protocole RoCE. RoCE est pris en charge sur les commutateurs de datacenter standard, et des améliorations supplémentaires ont été apportées pour optimiser les performances, telles que la commutation Ethernet de base élevée, la commutation Cut-Through, l'équilibrage de charge et des liaisons à bande passante plus élevée allant jusqu'à 800 GbE (4 x 200 GbE).

Les tests initiaux des LLM de GenAI peuvent fournir un aperçu précoce des avantages que GenAI peut apporter à l'entreprise et aider à élaborer une stratégie de LLM de GenAI, ainsi que sur les types d'infrastructure nécessaires. Essentiellement, la pile logicielle répond aux exigences en matière de semi-conducteurs pour la prochaine étape de l'évolution de l'IA générative d'entreprise. La compréhension de la pile logicielle vous aidera à déployer une infrastructure matérielle optimisée.

Rééquilibrage de l'infrastructure sur site par rapport à l'infrastructure hors site à mesure que les coûts des semi-conducteurs se stabilisent

À mesure que l'offre de processeurs graphiques de datacenter augmente, davantage de fournisseurs proposeront des processeurs graphiques de datacenter, davantage d'accélérateurs d'IA deviendront disponibles et davantage de puissance de traitement GenAI sera disponible pour les déploiements sur site. En parallèle, les goulots d'étranglement chez les fournisseurs de services cloud disparaîtront et les coûts devraient se stabiliser. Une fois que cela se produira, dans trois à cinq ans, le rééquilibrage des charges applicatives d'IA générative entre l'infrastructure sur site et l'infrastructure Cloud permettra d'optimiser les CAPEX et les OPEX.

Conclusion

L'IA générative est la technologie révolutionnaire pour l'IA. Les entreprises doivent disposer d'une stratégie/d'un plan d'IA générative qui doit être lancé dès maintenant pour les charges applicatives de taille d'entreprise afin de poursuivre le processus d'intégration de cette technologie révolutionnaire dans les opérations d'entreprise.

La demande est forte, ce qui fait grimper le prix des composants et celui des prestataires de services Cloud. Dans le même temps, IDC prévoit que plus de 50 % des systèmes d'IA générative ne seront pas accélérés en 2024 ; Par conséquent, n'importe qui peut commencer à déployer son infrastructure GenAI avec des serveurs et une gestion de réseau standard. Des processeurs graphiques sont également disponibles, pour ceux qui en ont besoin. Il existe plusieurs options pour déployer une infrastructure d'IA, ainsi que plusieurs types d'IA générative, d'IA, d'apprentissage automatique et de jumeaux numériques qui profiteront à différentes entreprises de différentes manières.

« Le marché a continué à sous-estimer la croissance de l'IA générative, et IDC s'attend à une croissance robuste de l'infrastructure et des semi-conducteurs de l'IA générative », a déclaré Brandon Hoff, IDC.

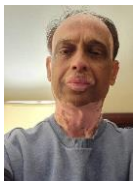
IDC prédit que les entreprises ramèneront les données du Cloud pour le traitement GenAI afin de réduire les coûts OPEX. Les entreprises commenceront leur développement et leurs tests avec GenAI sur du matériel de calcul et de gestion de réseau Ethernet standard et investiront au fur et à mesure qu'elles apprendront quels LLM fonctionnent pour leur entreprise et la valeur qu'ils peuvent extraire de leurs données propriétaires.

La création de l'infrastructure pour les tests LLM de GenAI sur des serveurs prêts à l'emploi et des réseaux Ethernet d'entreprise libérera la valeur de GenAI pour l'entreprise.

À propos des analystes

***Brandon Hoff**, directeur de recherche, Technologies habilitantes : gestion de réseau et communication*

Brandon Hoff dirige l'infrastructure de réseau et de communication d'IDC au sein de l'équipe Enabling Technologies d'IDC. M. Hoff couvre les tendances technologiques, les charges de travail, les produits, les fournisseurs, la chaîne d'approvisionnement et les stratégies d'adoption par les utilisateurs finaux dans l'informatique d'entreprise et les centres de données des fournisseurs de services Web, cloud et de télécommunications.



***Vijay Bhagavath**, Vice-président de la recherche, Cloud et réseaux de datacenter*

Vijay Bhagavath fournit un leadership éclairé et des informations pragmatiques sur les marchés et les technologies de gestion de réseau du cloud et des centres de données. Vijay a une compréhension approfondie du marché global des réseaux, des technologies, des feuilles de route des produits, de la différenciation concurrentielle et des stratégies de déploiement, ce qui lui permet de fournir des commentaires et des conseils perspicaces aux fournisseurs, aux fournisseurs de cloud, aux acheteurs informatiques d'entreprise et aux praticiens.

MESSAGE DU COMMANDITAIRE

Invitez l'IA dans vos données

Dell Technologies accélère votre transition, de la phase d'évaluation à la réalisation concrète, en vous permettant d'exploiter des technologies innovantes, une offre complète de services professionnels et un réseau étendu de partenaires.

- Simplifié. Accélérez les résultats en combinant des conseils stratégiques et des feuilles de route avec des solutions éprouvées et validées.
- Sur mesure. Tirez le meilleur parti de vos données avec une infrastructure conçue pour répondre aux besoins de votre entreprise.
- Fiable. Construisez votre avenir d'IA sur une base sécurisée, en protégeant vos données et votre propriété intellectuelle.

Offrez les meilleures performances d'IA et simplifiez l'approvisionnement, le déploiement et la gestion de l'infrastructure d'IA conçue pour l'ère de l'IA générative, avec la technologie, l'innovation et les avantages de Dell Technologies pour fournir des résultats plus intelligents et plus rapides.

Pour en savoir plus, voir www.dell.com/AI.

IDC Custom Solutions

Le contenu de ce document a été adapté à partir de l'étude IDC existante publiée sur www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, États-Unis
Tél. : 508.872.8200
Fax : 508.935.4015
Twitter : @IDC
idc-insights-community.com
www.idc.com