White Paper

Why Developing and Deploying AI Technology on Workstations Makes Sense

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

OPINION D'IDC

L'IA a pris son envol et devient un facteur de différenciation clé dans tous les secteurs d'activité. De plus, le matériel requis pour l'exécuter évolue rapidement. Le secteur des technologies est souvent très concentré sur l'augmentation exponentielle de la taille des modèles IA les plus avancés. Les discussions portent sur des dizaines de milliards de paramètres, sur la réduction de la précision, le développement de la mémoire, les besoins de type HPC (calcul haute performance) pour l'entraînement et l'inférence de l'IA, ainsi que sur les racks de serveurs accélérés. En réalité, l'informatique IA est rarement exécutée à une telle échelle, surtout dans les entreprises.

Aujourd'hui, bon nombre d'entre elles travaillent dur sur des initiatives IA, dont l'IA générative, qui ne nécessitent pas l'utilisation d'un superordinateur. En effet, une grande partie du développement de l'IA (et, de plus en plus, du déploiement de l'IA, notamment à la périphérie) s'effectue sur de puissantes stations de travail. Les stations de travail ont de nombreux avantages pour le développement et le déploiement de l'IA. Elles évitent au scientifique ou au développeur IA d'avoir à négocier du temps de serveur. Elles permettent de disposer de l'accélération fournie par les processeurs graphiques, encore difficilement disponibles sur les serveurs des datacenters. Beaucoup plus abordables que les serveurs, elles représentent une dépense ponctuelle moindre comparé à une instance Cloud pour laquelle la facture peut rapidement augmenter. De plus, elles permettent de gagner en sérénité, car les données sensibles sont stockées sur site, de manière sécurisée. Ainsi, le scientifique ou le développeur n'a plus à s'inquiéter de l'accroissement des coûts lorsqu'il procède à de simples expériences sur des modèles IA.

IDC constate que les déploiements IA augmentent plus rapidement à la périphérie que sur site ou dans le Cloud. Là aussi, les stations de travail jouent un rôle de plus en plus vital en tant que plateformes d'inférence IA. Elles ne requièrent souvent même pas de processeur graphique et sont capables d'effectuer l'inférence sur des processeurs optimisés par des logiciels. Les cas d'utilisation de l'inférence IA à la périphérie, sur des stations de travail, se multiplient rapidement et comprennent les AIOps, la réaction aux sinistres, la radiologie, l'exploration pétrolière et gazière, la gestion foncière, la télésanté, la gestion du trafic, la surveillance des usines de fabrication et les drones.

Ce livre blanc analyse le rôle croissant que les stations de travail jouent dans le développement et le déploiement de l'IA, et présente rapidement la gamme de stations de travail Dell pour l'IA.

July 2023, IDC n° US51017223

PRESENTATION DE LA SITUATION

L'explosion de l'IA et son impact sur l'infrastructure

Le nombre de projets IA entrepris par les organisations dans le monde augmente rapidement. Déjà, dans tous les secteurs d'activité, de nombreuses tâches sont effectuées par des logiciels partiellement ou totalement pilotés par un modèle IA. IDC procède au suivi de l'IA à de nombreux niveaux, et l'un des indicateurs intéressants à prendre en compte est le montant que les entreprises et prestataires de services Cloud consacreront aux serveurs pour développer et exécuter l'IA, selon les prévisions. D'ici 2026, ce montant atteindra 34,6 milliards de dollars, soit près de 22 % des dépenses mondiales totales consacrées aux serveurs.

Toutefois, l'IA ne se limite pas aux serveurs. Une grande partie de la préparation, du développement, du prototypage et, de plus en plus, du *déploiement* s'effectue sur des stations de travail. Alors que les organisations, petites et grandes, découvrent qu'elles peuvent saisir de nouvelles opportunités commerciales en ajoutant des fonctionnalités IA à leurs applications, les expériences impliquant des modèles IA ont explosé, et les stations de travail robustes sont idéales dans ces cas-là, de par leur disponibilité immédiate et leur proximité avec les données.

Pourquoi l'IA s'est-elle répandue si soudainement, alors que des algorithmes IA sont déployés depuis des décennies ? C'est principalement parce que deux conditions exigées pour alimenter un type particulièrement efficace d'algorithme IA, le réseau neuronal, ont pu être réunies ces quelques dernières années : la grande disponibilité de types de données à la fois variés, économiques et volumineux, comme les données non structurées et semi-structurées, et l'augmentation du calcul linéaire avec un modèle parallèle pour traiter ces réseaux neuronaux dans un délai acceptable. Depuis que ces deux conditions de base sont satisfaites, les scientifiques des données ont fait d'incroyables progrès dans le développement de réseaux neuronaux qui apprennent automatiquement comment exécuter des tâches de plus en plus impressionnantes. Si l'apprentissage automatique (ML) traditionnel reste utile pour les données textuelles et numériques, le Deep Learning (DL) est plus efficace pour la vidéo, l'audio, les langues, etc.

Les modèles d'apprentissage automatique traditionnels peuvent généralement être développés sur les processeurs d'une station de travail, qui comptent plusieurs dizaines de cœurs tout au plus, mais les réseaux neuronaux requièrent des coprocesseurs dont le traitement s'effectue en parallèle sur des milliers de cœurs. La principale raison est la suivante : en apprentissage automatique, l'extraction et la classification de caractéristiques est un processus manuel, tandis qu'en Deep Learning, ces opérations sont automatisées et exigent l'entraînement du modèle par des répétitions constantes à l'aide de grands jeux de données. Actuellement, le coprocesseur le plus courant est le processeur graphique, mais de nouveaux processeurs propres à l'IA, développés par des start-up, sont en train d'apparaître également. Ce type d'accélération, utilisant un coprocesseur indépendant pour le traitement en parallèle, a révolutionné les marchés des serveurs et des stations de travail, et a ouvert la voie à ce qu'IDC appelle le calcul massivement parallèle.

En 2022, les serveurs accélérés représentaient un marché mondial de 21,8 milliards de dollars. Celui-ci devrait croître jusqu'à 43,4 milliards de dollars d'ici 2026, sachant que 57 % de ce montant total sont imputables aux serveurs accélérés servant à exécuter l'IA. Dans le même temps, le nombre de processeurs graphiques indépendants vendus pour être utilisés dans des stations de travail est passé à 6,4 millions en 2022. IDC estime que le marché des stations de travail utilisées à des fins scientifiques ou d'ingénierie logicielle, de plus en plus encouragées par le développement de l'IA, augmentera pour atteindre près de 2 milliards de dollars d'ici 2026.

Étapes du développement de l'IA

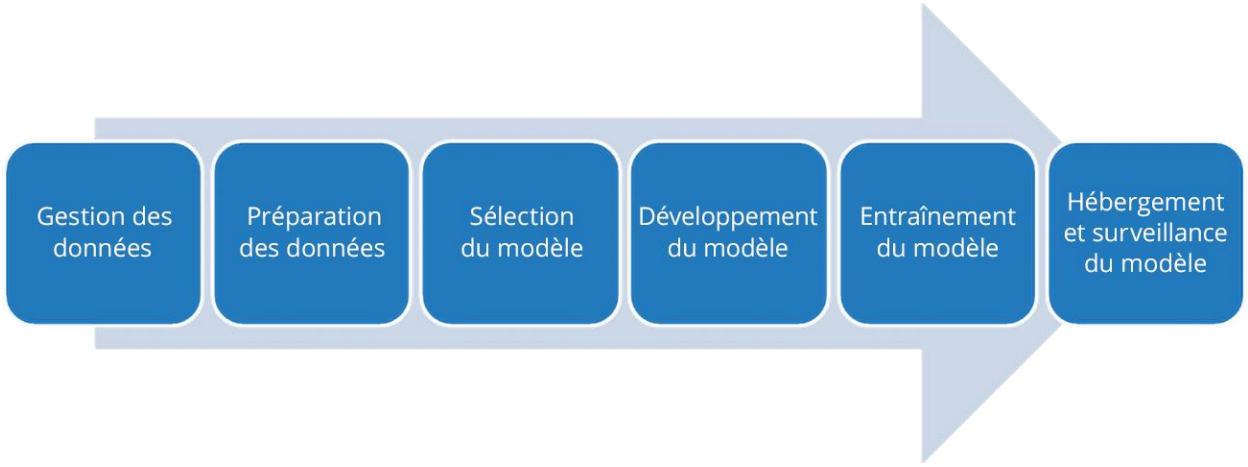
Comme nous l'avons précédemment mentionné, le développement des types et volumes de données ainsi que de nouvelles approches de calcul a permis de faire fonctionner les réseaux neuronaux. La première variable de cette équation (les types et volumes de données) n'est pas des moindres. Selon certaines sources, 80 % des efforts d'une initiative IA de Deep Learning sont consacrés à la gestion et à la préparation des données. Les données doivent être ingérées, gérées et préparées avant que la conception et l'entraînement de modèles ne puissent commencer. Selon IDC, les étapes du développement de l'IA sont les suivantes (voir figure 1) :

- **Gestion des données :** identification et gestion des données pertinentes pour le modèle IA, à partir des immenses volumes de données du datacenter, de la périphérie et du Cloud qu'une organisation ingère, génère et/ou acquiert (ces données pouvant être de tous types, provenir d'événements ou du streaming et, pour la plupart, requérir une certaine forme de gouvernance).
- **Préparation des données :** stockage des données (en mode fichier, bloc ou objet) dans un entrepôt de données ou un Data Lake, nettoyage de ces dernières, vérification de leur exhaustivité et de leur haute qualité, puis conversion de ces données dans un format exploitable par le modèle IA, par exemple avec Spark ou des outils comme Pandas.
- **Sélection du modèle :** décision consistant à choisir le modèle qui exécute de manière optimale la tâche IA pour laquelle il est programmé, en matière de taux d'erreur et/ou de performances.
- **Développement du modèle :** conception du modèle IA à l'aide de cadres tels que XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn et H2O.
- **Entraînement du modèle :** entraînement du modèle sur l'infrastructure de calcul avec suffisamment de cœurs de processeur et/ou coprocesseur pour la parallélisation (impliquant également de plus en plus la capacité à expliquer, valider et documenter les décisions relatives à un modèle afin d'assurer l'équité, la responsabilité et la transparence). (Cette étape comprend le prototypage, c'est-à-dire le fait de tester le modèle entraîné en effectuant l'inférence dessus.)
- **Hébergement et surveillance du modèle :** déploiement du modèle dans un environnement de production afin qu'il exécute la tâche pour laquelle il a été conçu, généralement appelée « inférence IA », et surveillance de ses performances.

Les stations de travail peuvent jouer un rôle important dans toutes ces étapes, en combinaison avec le datacenter, le Cloud ou l'infrastructure de périphérie.

FIGURE 1

Étapes du développement de l'IA



Source : IDC, 2023.

DEVELOPPEMENT DE MODELES IA SUR DES STATIONS DE TRAVAIL

Comparatif entre stations de travail et ordinateurs personnels

Généralement, tout le monde comprend bien que les ordinateurs personnels (PC) ne sont pas suffisamment puissants pour le développement de l'IA. Les scientifiques des données et les développeurs IA sont habituellement impliqués dans des projets stratégiques pour leur organisation. Une productivité optimale est donc de la plus haute importance. Les stations de travail ont tendance à fonctionner de manière plus prévisible que les PC, car elles sont généralement équipées de composants qui offrent des performances supérieures et optimisées pour les logiciels qui s'exécutent dessus.

Ces composants sont notamment les suivants :

- **Processeurs de haute qualité :** par exemple, processeurs Intel Xeon Scalable.
- **Puissants processeurs graphiques :** par exemple, processeurs graphiques professionnels NVIDIA RTX, dont le modèle NVIDIA RTX 6000 Ada.
- **Davantage de stockage :** certaines stations de travail peuvent fournir jusqu'à 60 To de stockage, et les vitesses d'E/S tendent à être considérablement supérieures à celles des PC.
- **Davantage de mémoire :** les stations de travail disposent désormais d'une mémoire allant jusqu'à 6 To.
- **Refroidissement :** les composants hautes performances génèrent beaucoup de chaleur, et les scientifiques des données ont besoin d'une station de travail équipée d'un système de refroidissement approprié pour éviter toute surchauffe et garantir des performances optimales.
- **Carte d'interface réseau (NIC) :** les scientifiques des données qui travaillent sur de vastes jeux de données stockés sur des serveurs distants ont besoin d'une carte NIC haut débit pour pouvoir transférer les données rapidement et efficacement.

- **Affichage :** un affichage de haute qualité est important pour les tâches de visualisation des données, et les scientifiques des données doivent chercher un grand écran haute résolution offrant des couleurs précises.
- **Mémoire ECC (Error Correction Code) :** la mémoire ECC détecte et corrige les types les plus courants de corruptions des données internes, ce qui évite que des écrans bleus n'apparaissent au cours d'un long entraînement de l'IA, que ce soit à cause d'une erreur importante (mauvais bit) ou d'une erreur secondaire (bit permuté causant de mauvaises valeurs) ; la mémoire ECC garantit également la précision des résultats, ce qui est crucial pour les tâches vitales comme celles des services de santé.
- **Puces spécialisées :** par exemple, les unités de traitement de la vision Intel Movidius, qui sont des coprocesseurs de traitement parallèle pour les applications de vision par ordinateur et d'IA en périphérie, utilisées dans les secteurs de la vente au détail, de la sécurité et de l'automatisation industrielle. Les FPGA (Field-Programmable Gate Array, réseau de portes programmables in situ) sont aussi utilisés dans les stations de travail, par exemple, pour les applications financières.
- **Logiciel d'optimisation :** par exemple, oneAPI, qui est le modèle de programmation standardisé d'Intel pour simplifier le développement et le déploiement de charges applicatives axées sur les données entre des processeurs, processeurs graphiques, FPGA et d'autres accélérateurs, ou CUDA, qui est la plateforme de traitement parallèle et l'interface de programmation d'application de NVIDIA pour exécuter des charges applicatives d'ordre général sur des processeurs graphiques.

Comparatif entre processeurs et processeurs graphiques pour l'IA

Les stations de travail peuvent être utilisées à diverses étapes du développement de l'IA, et elles sont généralement équipées pour offrir une variété de fonctionnalités. Bien que l'accent soit mis sur les processeurs graphiques pour le traitement parallèle, les processeurs jouent un rôle essentiel lors du développement d'un modèle IA sur une station de travail. Tout comme les processeurs graphiques, les processeurs peuvent être utilisés pour la manipulation des données et, bien entendu, pour le développement de modèles ML traditionnels. Les processeurs sont également utilisés pour l'exploration des données, c'est-à-dire le processus consistant à se servir des représentations visuelles d'un jeu de données pour comprendre les caractéristiques de ces dernières.

Dans l'entraînement DL, le rôle des processeurs hôtes est quelque peu réduit car les processeurs graphiques prennent la relève au cours du processus d'entraînement lui-même. Pourtant, même dans ces cas, les processeurs continuent de servir en tant que couche de traitement pour les logiciels stratégiques, comme le système d'exploitation ou CUDA, et pour l'orchestration des processus entre les processeurs graphiques ou d'autres puces. En outre, les processeurs jouent de plus en plus le nouveau rôle de moteurs d'inférence IA dans les cas où les stations de travail sont utilisées pour exécuter un modèle IA en production. IDC s'attend à ce que, d'ici 2024, les dépenses consacrées à l'infrastructure pour l'inférence IA dépassent celles destinées à l'infrastructure IA pour l'entraînement IA, et à ce qu'une importante partie (39 %) de cette inférence soit effectuée sur les processeurs hôtes.

Stations de travail et serveurs : une relation symbiotique

La plupart des organisations font preuve d'un grand pragmatisme lorsqu'elles déploient une station de travail, un serveur sur site, une instance Cloud ou une combinaison des trois pour le développement de l'IA. Il existe une relation symbiotique entre les stations de travail, les serveurs et les instances Cloud pour les différentes étapes de développement d'un projet IA.

L'avantage des stations de travail par rapport aux serveurs de datacenter est qu'elles permettent aux scientifiques des données de travailler où ils le souhaitent, ce qui est important dans le cadre de la pandémie actuelle, mais également en temps normal. Ils peuvent également mener librement leurs expériences sur leurs modèles IA, en itérant autant de fois qu'ils le jugent nécessaire. En effet, la puissance des stations de travail modernes, équipées de processeurs graphiques performants, permet souvent davantage d'interactions au cours du processus itératif. Ils obtiennent ainsi des commentaires et des résultats instantanés, sans avoir à demander l'accès aux serveurs ou à se confronter à d'autres restrictions. Les stations de travail leur offrent également la flexibilité nécessaire pour rapprocher le calcul des données, plutôt que l'inverse, ce qui économise de la bande passante, réduit la congestion du réseau et augmente le débit. Autre avantage : les stations de travail peuvent être configurées pour répondre à différents besoins, que ce soit pour des tâches ML traditionnelles ou pour un travail de DL plus intensif.

De plus, même si le marché des serveurs accélérés augmente considérablement, ces derniers ne sont pas encore largement répandus dans les datacenters d'entreprise. Au moment où nous rédigeons ce livre blanc, en moyenne, 4 % des serveurs des datacenters d'entreprise sont accélérés, ce qui signifie que bon nombre des organisations n'ont pas la possibilité de développer ni d'exécuter l'IA sur des processeurs graphiques sur site facilement disponibles. C'est pour cette raison, également, que les stations de travail accélérées sont une alternative utile pour le développement de l'IA.

Les stations de travail hautement accélérées sont désormais assez puissantes pour pouvoir réaliser l'entraînement DL, tant que le modèle IA n'est pas excessivement volumineux, ce qui évite d'avoir à faire appel à des serveurs. Et les modèles entraînés sur des stations de travail équipées de processeurs graphiques peuvent être déployés soit sur des stations de travail, soit sur des serveurs sans processeur graphique, en utilisant les fonctionnalités d'inférence des processeurs. Les technologies logicielles comme DL Boost et oneAPI d'Intel peuvent optimiser l'inférence IA sur le processeur, ce qui permet aux serveurs non accélérés déjà déployés dans les datacenters de prendre en charge les applications IA.

Comparatif entre stations de travail et Cloud

Le Cloud Computing a révolutionné la façon dont les organisations réfléchissent à l'infrastructure, aux données et aux applications. En promettant une évolutivité pratiquement illimitée, le Cloud donne la possibilité aux développeurs de provisionner les ressources à la demande, ce qui peut accélérer le rythme de l'innovation en réduisant les contraintes. À première vue, le Cloud apparaît comme le parfait paradigme pour le développement de l'IA.

Toutefois, ce n'est pas toujours le cas. En réalité, les recherches d'IDC ont montré que les organisations rapatrient de plus en plus certaines charges applicatives du Cloud public vers l'infrastructure sur site. Plusieurs facteurs expliquent ce phénomène :

- **Disponibilité du Cloud :** toute personne ayant utilisé des services Cloud a déjà subi une panne, que cette dernière ait été due à des problèmes chez le fournisseur de Cloud ou à un souci de connectivité réseau entre le datacenter hyperscale et l'utilisateur final. Dans ces situations-là, les utilisateurs sont à la merci du prestataire de services qui doit résoudre le problème, tandis que la productivité est à l'arrêt.
- **Sécurité et conformité :** dans de nombreux secteurs, les règles de gouvernance des entreprises dictent où les données peuvent être communiquées et stockées, ce qui limite l'utilisation des services Cloud. Les réglementations gouvernementales comme le RGPD en Europe et le California Consumer Privacy Act imposent également des règles de souveraineté des données.

- **Coût :** il est courant que les organisations sous-estiment la rapidité avec laquelle les frais des services Cloud peuvent augmenter, surtout pour les charges applicatives qui requièrent des capacités de calcul haute performance et de grandes quantités de stockage. Le concept économique du Cloud consiste à mesurer tous les types de ressources consommées, y compris le retrait des données pour les renvoyer vers l'infrastructure sur site.
- **Pression liée aux essais/erreurs :** la plupart des initiatives IA commencent par une certaine quantité d'expériences. Les modèles qui échouent font donc partie intégrante du processus de développement. Toutefois, dans le cadre de ce processus, lorsque la facture Cloud augmente et qu'aucun résultat exploitable ne vient, les scientifiques et développeurs IA subissent une forte pression psychologique.

Les stations de travail peuvent supprimer ces inconvénients, tout en permettant d'utiliser des technologies Cloud natives comme les architectures basées sur des microservices et l'automatisation pilotée par API. Cela permet d'obtenir certains avantages tels que ceux présentés lors du comparatif entre stations de travail et serveurs de datacenter :

- **Travail en tout lieu :** en supprimant la dépendance au Cloud public, des scénarios déconnectés sont désormais possibles. Bon nombre d'environnements haute sécurité sont isolés par air gap des réseaux publics, et seules des stations de travail IA peuvent être utilisées dans ce cas. Les ressources locales permettent également de se passer d'une connectivité réseau onéreuse.
- **Localité des données :** la prolifération des appareils IoT et autres équipements connectés contribue à la croissance exponentielle des données à la périphérie. Dans de nombreuses situations, il est judicieux de colocaliser les ressources de calcul avec une station de travail dédiée. Cela permet également de satisfaire à bon nombre d'exigences de conformité en limitant les mouvements des données.
- **Liberté d'expérimentation :** l'entraînement et l'optimisation de modèles IA sont des processus itératifs, qui impliquent souvent des essais et des erreurs. Les développeurs ont besoin de pouvoir procéder à des expériences en toute liberté, sans avoir à faire de compromis à cause d'éventuels frais de services supplémentaires. Les stations de travail fournissent aussi davantage de flexibilité pour un outillage sur mesure.

Concernant ce dernier point, comparer le prix d'une station de travail à celui d'un déploiement Cloud est relativement simple, car la plupart des prestataires de services Cloud fournissent une estimation de coûts instantanée pour toute configuration qu'un utilisateur final souhaiterait déployer. Par exemple, le coût d'une seule machine virtuelle (VM) classique avec un NVIDIA T4 et une instance de stockage SSD de 375 Gio utilisée huit heures par jour et cinq jours par semaine est de 140 $ chez un grand fournisseur de Cloud. Si vous doublez le nombre de VM, de T4 et de SSD, le coût augmente jusqu'à 365 $ par mois. Si vous restez à deux VM, mais doublez les T4 pour en avoir quatre, doublez le stockage pour atteindre 4 x 375 Gio, et que vous réalisez un entraînement à plein temps sur l'environnement, le coût bondit jusqu'à 2 700 $ par mois. Il est donc juste de dire que les coûts du Cloud pour le développement de l'IA peuvent facilement atteindre des dizaines de milliers de dollars par an, ce qui représente considérablement plus que l'amortissement annuel d'une station de travail haut de gamme.

PROTOTYPAGE IA SUR LES STATIONS DE TRAVAIL

Comparées aux serveurs sur site et au Cloud, les stations de travail fournissent un net avantage en matière de prototypage de modèles IA. Les serveurs du datacenter peuvent être totalement utilisés ou trop stratégiques pour servir au prototypage et aux tests IA, et comme nous l'avons indiqué précédemment, les instances Cloud peuvent rapidement générer des dépassements de frais lorsqu'elles sont fortement utilisées comme environnement de test. Les stations de travail, elles, évitent au scientifique ou au développeur IA d'avoir à négocier du temps de serveur ou à subir des remarques continuelles sur l'augmentation des factures Cloud durant l'étape de prototypage. Leur faible coût unique offre une liberté totale pour procéder au prototypage en tout lieu et à tout moment, sans frais supplémentaires.

DEPLOIEMENT DE MODELES IA SUR DES STATIONS DE TRAVAIL

Alors que le développement de modèles IA sur une station de travail constitue une stratégie courante depuis des années, IDC observe de plus en plus de cas d'utilisation consistant à *déployer* un modèle IA sur une station de travail, généralement à la périphérie. Autrement dit, il s'agit de mettre le modèle IA en production sur la station de travail en lui faisant exécuter l'inférence sur le modèle IA. L'utilisation de la périphérie comme site de déploiement de l'IA pour les serveurs prend rapidement de l'ampleur. Elle a plus que triplé entre 2020 et 2024, en termes de dépenses matérielles annuelles. Et les stations de travail ne sont pas loin derrière, car les utilisateurs finaux découvrent leurs avantages à la périphérie.

IDC définit la périphérie comme un paradigme informatique distribué comprenant le déploiement d'infrastructures et d'applications hors d'un Cloud centralisé et de datacenters sur site, aussi proche que nécessaire du lieu de génération et de consommation des données. La périphérie inclut donc les bureaux distants et filiales ainsi que les sites propres à un secteur comme les usines, entrepôts, hôpitaux et magasins de vente au détail.

Les charges applicatives utilisant de grands volumes de données et exigeantes en ressources de calcul sont de plus en plus déployées sur site ou à la périphérie. Cela permet d'éviter les inconvénients inhérents aux Clouds publics tels que le temps nécessaire pour charger de grands jeux de données et les coûts variables de l'entraînement IA, notamment dans les cas qui requièrent de nombreuses expériences de science des données.

Les recherches d'IDC montrent que les scénarios de déploiement de l'IA à la périphérie se multiplient rapidement. Les organisations ont investi 2,9 milliards de dollars dans le calcul IA en périphérie en 2023 et y consacreront 6,9 milliards de dollars en 2026 (voir *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC n° US49671722, septembre 2022). De plus, les déploiements de charges applicatives HPC (par exemple, ingénierie et technique) à la périphérie prennent de l'ampleur. Les entreprises investissent actuellement près de 1 milliard de dollars dans ces charges applicatives à la périphérie et y consacreront 2,4 milliards de dollars d'ici 2027 (voir *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC n° US50525123, avril 2023). Dans ces domaines, il est judicieux de déployer une station de travail IA.

Lorsque l'on déploie un modèle IA sur une station de travail à la périphérie, il n'est pas toujours nécessaire d'utiliser des processeurs graphiques haut de gamme, comme c'est le cas pour le développement de l'IA. Des processeurs graphiques plus légers peuvent effectuer l'inférence IA, et dans un bon nombre de cas, les processeurs graphiques ne sont tout simplement pas nécessaires. Dans ces cas-là, les processeurs peuvent parfaitement effectuer la tâche d'inférence, notamment lorsqu'ils sont utilisés avec des optimisations comme Intel DL Boost, un ensemble d'instructions qui se trouve sur les microprocesseurs Intel et qui est conçu pour accélérer les charges applicatives IA, dont l'inférence IA. Avec Intel DL Boost, Intel déclare avoir observé un débit d'inférence INT8 en temps réel 1,45 fois supérieur avec le processeur Intel Xeon Scalable de 4e génération qui prend en charge Intel DL Boost par rapport à la génération précédente (BERT-Large SQuAD). Avec ce type d'optimisation, une station de travail peut être plus facilement déployée à la périphérie, où certains aspects comme l'alimentation, la mobilité et la gestion thermique requièrent une puissance électrique inférieure. Intel Movidius Myriad (M2) s'adapte bien dans cette enveloppe énergétique grâce à sa petite empreinte de 12 W.

Cas incitant à déployer l'IA sur des stations de travail

Plusieurs situations se prêtent naturellement au déploiement de l'IA sur des stations de travail déployées localement. Elles ont en commun de grands volumes de données de séries temporelles générées par des machines et de données non structurées comme les flux vidéo et les images. Il existe également des cas où les experts SME doivent compléter les modèles IA par une interprétation humaine.

Voici quelques exemples :

- AIOps : alors que les systèmes IT gagnent en taille et en complexité, il devient de plus en plus nécessaire de passer de la gestion réactive des incidents à la surveillance proactive. C'est particulièrement vrai lorsque l'infrastructure et les applications sont distribuées sur des sites de périphérie où il y a peu, voire pas de personnel technique. En modélisant une base de référence de performances normales, il est possible d'identifier les anomalies et d'automatiser les étapes correctives.
- Réaction aux incidents : en cas d'urgence, les premiers intervenants doivent rapidement évaluer la situation, localiser les équipements vitaux et déployer des ressources pour aider ceux qui en ont le plus besoin. Toutes ces opérations doivent souvent être effectuées dans un environnement sans connectivité réseau, ce qui nécessite une station de travail locale capable d'agréger les flux de données, d'inférer par rapport aux modèles IA et d'automatiser les communications vers le personnel clé.
- Radiologie : les progrès des technologies d'imagerie ont entraîné l'augmentation de la taille des données générées lors d'un scanner. Ces dernières doivent donc rester sur site pour être analysées dans les meilleurs délais. Les modèles IA entraînés à partir de millions d'exemples précédents peuvent identifier des structures et formes plus précisément que l'œil humain, ce qui augmente les taux de précision.
- Exploration pétrolière et gazière : les compagnies pétrolières et gazières de l'upstream utilisent une combinaison de données sismiques, de télémétrie et d'imagerie pour localiser des réserves de ressources naturelles, sélectionner les lieux de forage et optimiser les performances de l'équipement intervenant dans le processus de production. Cela nécessite souvent d'analyser des informations dans des zones où seule une communication onéreuse par satellite est possible.

- **Recherche contre le cancer et développement de médicaments :** les chercheurs hospitaliers et universitaires utilisent l'IA et le traitement du langage naturel pour aider les oncologues à trouver le traitement individualisé le plus efficace pour leurs patients atteints de cancer. Ils associent également l'apprentissage automatique à la vision par ordinateur pour permettre aux radiologues de mieux comprendre l'évolution des tumeurs des patients. Et ils utilisent des algorithmes pour mieux comprendre comment les cancers se développent et quels traitements fonctionnent le mieux contre eux.
- **Évaluations des déclarations de sinistre :** le traitement manuel des déclarations est à la fois fastidieux et sujet aux erreurs humaines. L'IA peut évaluer la validité des déclarations, ce qui réduit les coûts en permettant aux experts en assurance de se concentrer sur les cas qui requièrent davantage d'investigation. Cela accroît le débit global de l'opération, sans sacrifier la précision.
- **Télésanté :** l'IA améliore les taux de rétablissement des patients en adaptant les plans de traitement individuels d'après les signes vitaux fournis en temps réel par des appareils connectés portables. Ces informations sont combinées avec les dossiers historiques des patients et une base de connaissances de cas similaires. Ceci est particulièrement important dans les zones rurales qui dépendent davantage des services de santé distants.
- **Sécurité de la vente au détail (antivol) :** l'analytique en temps réel appliquée aux flux vidéo est utilisée pour prédire les comportements humains pouvant conduire à des activités criminelles. Elle implique généralement de raccorder plusieurs flux vidéo pour suivre les mouvements d'un individu dans un magasin. Étant donné qu'il faut rapidement identifier un événement matériel, il vaut mieux effectuer ce processus localement.
- **Gestion du trafic :** les organismes gouvernementaux responsables d'opérations de transport utilisent de plus en plus l'IA pour assurer la coordination des feux de circulation et de la signalisation numérique afin d'améliorer le flux des véhicules et de garantir la protection des citoyens. Cela nécessite une combinaison d'entrées, notamment de caméras vidéo et de télémétrie issue de capteurs situés sur les routes, afin d'optimiser les schémas de trafic.
- **Surveillance des usines de fabrication :** pour un responsable d'usine, il est crucial de garantir le temps d'activité des processus stratégiques et de respecter les calendriers de production. Cela implique d'assurer la maintenance prédictive des principaux équipements, d'automatiser la détection des défauts et de procéder à des optimisations à la fois à l'intérieur et à l'extérieur de la chaîne logistique du site. Dans ces domaines, l'IA peut aider les opérateurs humains à accroître les performances, tout en préservant les normes de sécurité.
- **Drones :** l'analyse automatisée d'images capturées par des drones permet de surveiller un large éventail de conditions à grande échelle, ce qui n'était pas possible auparavant. Cela est extrêmement utile pour l'inspection des infrastructures des fournisseurs de gaz et d'électricité, les enquêtes d'assurance, les missions de recherche et de sauvetage, l'agriculture de précision, ainsi que pour la protection des zones de pêche et des réserves naturelles.
- **Environnements de bureau quotidiens :** ces environnements sont de plus en plus améliorés par des outils de productivité basés sur l'IA comme Microsoft Copilot.
- **Énergies renouvelables :** les sites d'énergies renouvelables, comme les fermes éoliennes, solaires et les barrages hydroélectriques, requièrent une surveillance et une maintenance en temps réel ainsi que la collecte de données qui doivent être générées et analysées localement.

STATIONS DE TRAVAIL DELL POUR L'IA

Dell propose une large gamme de stations de travail pour différents niveaux de développement et/ou d'implémentation de l'IA. Ces dernières sont regroupées sous la marque Data Science Workstation (DSW). Cette section exposera rapidement les spécifications de ces modèles, puis présentera de nombreuses typologies d'utilisateurs/applications IA (par exemple, scientifiques des données) ainsi que les avantages de la technologie Dell DSW. Ces stations de travail pour la science des données, prêtes pour l'IA, ont été spécialement conçues pour les scientifiques des données. Les dernières stations de travail Precision pour la science des données utilisent des fonctionnalités IA pour régler précisément les appareils afin d'optimiser les performances des applications que les scientifiques des données utilisent le plus. Cela leur permet de réaliser leurs tâches les plus importantes plus rapidement. En outre, les stations de travail Dell Precision sont testées et certifiées par des éditeurs de logiciels indépendants (ISV) pour garantir qu'elles prennent en charge les applications hautes performances requises par les clients Dell pour effectuer leurs tâches quotidiennes.

Facteurs de différenciation des stations de travail Dell

Les stations de travail Dell Precision, accélérées par des processeurs graphiques NVIDIA RTX, sont conçues pour fournir une grande évolutivité et de hautes performances afin de prendre en charge les initiatives d'analytique et d'IA d'une organisation. Dell Technologies fournit des solutions matérielles complètes, optimisées pour exécuter les derniers logiciels IA du secteur :

- **Configuration matérielle robuste :** les stations de travail Dell Precision proposent diverses configurations matérielles puissantes, comprenant des processeurs multicœurs, une grande quantité de mémoire RAM et plusieurs options de processeur graphique. Ces composants fournissent les ressources de calcul nécessaires pour les tâches IA, ce qui permet un entraînement et une inférence efficaces.
- **Évolutivité et possibilités de personnalisation :** les stations de travail Dell Precision sont évolutives et personnalisables, ce qui permet aux utilisateurs d'adapter la configuration matérielle à leurs exigences IA spécifiques. Cette flexibilité garantit que la station de travail peut être optimisée pour répondre aux besoins particulier des charges applicatives IA.
- **Certification et optimisation :** Dell collabore avec NVIDIA pour certifier la compatibilité et les performances des stations de travail Precision avec les processeurs graphiques NVIDIA RTX, dont les cartes NVIDIA RTX 6000 Ada. Cette certification garantit une parfaite intégration et des performances optimisées lors de l'utilisation de stations de travail Dell Precision avec des processeurs graphiques NVIDIA RTX pour des tâches IA.
- **Puissante capacité de traitement :** les stations de travail Dell Precision, équipées de processeurs Intel, fournissent la puissance de calcul requise pour les tâches IA. Avec leurs processeurs multicœurs et des vitesses d'horloge élevées, ces stations de travail fournissent les performances requises pour l'entraînement et l'inférence dans le cadre de workflows IA.
- **Prise en charge par des logiciels et outils :** les stations de travail Dell Precision sont dotées de logiciels et d'outils qui prennent en charge le développement et le déploiement de l'IA. Cela inclut des piles de logiciels optimisées, des cadres IA et des bibliothèques tirant parti des processeurs graphiques NVIDIA RTX. Ainsi, les utilisateurs peuvent plus facilement se lancer dans des projets IA.

De plus, les technologies présentées dans les sections qui suivent sont d'autres facteurs de différenciation clés des stations de travail Dell.

Reliable Memory Technology

Outre l'ECC, Dell fournit une solution nommée Reliable Memory Technology Pro (RMT Pro), conçue pour aider à maximiser le temps d'activité. Elle fonctionne conjointement avec la mémoire à code de correction d'erreur (ECC) pour détecter et corriger les erreurs de mémoire en temps réel. Selon Dell, RMT Pro élimine pratiquement toutes les erreurs de mémoire en empêchant tout nouvel accès aux zones incorrectes, même si le module DIMM reste totalement utilisable. Après un redémarrage système, RMT Pro isolera la zone de mémoire défectueuse et la cachera au système d'exploitation. Cela évitera aux scientifiques des données et aux développeurs IA d'être constamment confrontés à des pannes dues au fait qu'une zone de mémoire défectueuse soit restée adressable, ce qui représente un immense gain de productivité.

Dell Optimizer for Precision

Dell ajoute également Dell Optimizer for Precision à la plupart de ses stations de travail. Cette solution ajuste automatiquement les paramètres système pour que la station de travail exécute diverses applications professionnelles populaires le plus rapidement possible. Cela améliore la productivité du scientifique des données ou du développeur. Cet outil crée également des rapports de performances en temps réel pour l'équipe IT, au sujet de l'utilisation du processeur, du stockage, de la mémoire et de la carte graphique. DOP ne s'exécute pas encore sur Linux. Il est donc surtout utile pour le déploiement de l'IA, car le développement de l'IA tend à être effectué avec des logiciels Open Source basés sur Linux. Dell Optimizer for Precision fournit également les fonctionnalités ExpressSign-in, ExpressCharge (sur les modèles mobiles), Intelligent Audio, ainsi que des outils de reporting et d'analytique pour aider à régler précisément la station de travail.

DEFIS/OPPORTUNITES

Pour les entreprises

IDC constate que le marché de l'IA est scindé en deux. D'un côté, certaines entreprises déploient des stratégies de données pour rester concurrentielles, y compris en se lançant dans d'importants projets d'IA. À titre d'exemple, on leur présente des homologues qui ont fourni un travail extraordinaire en utilisant des offres d'infrastructure IA d'entreprise figurant véritablement au top 100 des superordinateurs. D'un autre côté, certaines entreprises sont confrontées à la réalité quotidienne de petites initiatives IA testées sur les serveurs disponibles du datacenter ou dans le Cloud, qui riment souvent avec des budgets insuffisants et du matériel sous-performant.

Pour de nombreuses entreprises, le premier scénario n'est tout simplement pas envisageable et le second malheureusement trop réel. Pour ces entreprises, la difficulté est de fournir aux scientifiques des données et/ou aux développeurs les bons outils pour procéder à l'entraînement de l'IA, au bon moment, sans dépenser énormément d'argent dans des instances Cloud ou des serveurs de datacenter accélérés par des processeurs graphiques. IDC pense que ces entreprises ont tout intérêt à fournir à leurs scientifiques et développeurs de puissantes stations de travail accélérées par des processeurs graphiques.

Pour Dell

Les entreprises pensent souvent à tort que le développement et le déploiement de l'IA requièrent du matériel serveur accéléré et onéreux, souvent même dans un cluster. Cela peut être vrai pour les plus grands algorithmes IA impliquant des milliards de paramètres, mais la plupart des entreprises ne développent pas de tels algorithmes. Elles cherchent à mener des initiatives IA utiles, efficaces et gérables, mais bon nombre d'entre elles ne se rendent pas compte que ces modèles IA à échelle normale peuvent être développés, et déployés, sur des stations de travail. Dell doit donc combattre les idées reçues et informer le marché des possibilités offertes par sa gamme de stations de travail.

Parallèlement, Dell doit s'assurer que ses stations de travail sont capables de fournir les performances nécessaires et éviter qu'elles ne se transforment en goulets d'étranglement avec le temps. Autrement dit, Dell doit innover rapidement et en permanence pour ne jamais décevoir les utilisateurs finaux qui utilisent les stations de travail de manière appropriée (c'est-à-dire qui n'essaient pas d'exécuter des algorithmes comptant plusieurs milliards de paramètres). Quant aux clients dont les initiatives prennent soudain très rapidement de l'ampleur et dont les algorithmes deviennent immenses, ils peuvent facilement évoluer des stations de travail vers la gamme de serveurs Dell destinés à l'IA. C'est l'occasion pour Dell de proposer à chaque client la solution qui lui correspond, quelle que soit la taille de l'initiative IA sur laquelle il travaille.

CONCLUSION

IDC pense que les stations de travail sont actuellement sous-estimées, alors qu'elles sont suffisamment puissantes pour prendre en charge le développement et le déploiement de l'IA dans de nombreux cas d'utilisation. Elles constituent pour les scientifiques et développeurs IA de puissantes plateformes accélérées par des processeurs graphiques, impliquent moins de CAPEX que les serveurs, considérablement moins d'OPEX que les instances Cloud et offrent beaucoup plus de liberté pour tester les modèles IA. Les entreprises menant des initiatives IA qui ne requièrent pas d'algorithmes avec des milliards de paramètres (comme la plupart) devraient envisager d'équiper leurs collaborateurs IA de stations de travail pour faciliter le développement de l'IA et le déploiement à la périphérie.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com