MÉDIAS ET DIVERTISSEMENT
JAPON

Développement de l'IA générative en japonais et transformation des services de publicité numérique

CyberAgent, Inc. utilise des serveurs Dell PowerEdge XE9680 équipés de huit processeurs graphiques NVIDIA® H100 Tensor Core pour accélérer l'IA générative et améliorer l'efficacité des publicités.

Besoins de l'entreprise

Depuis 2016, CyberAgent, Inc. s'investit activement dans la recherche et le développement de l'IA, ainsi que dans l'intégration de cette technologie dans ses activités publicitaires. L'entreprise avait besoin de fournir à ses équipes un accès rapide et abordable à des serveurs sur site extrêmement fiables, équipés des processeurs graphiques NVIDIA les plus avancés pour soutenir ses efforts en matière de développement de l'IA générative.

Résultats commerciaux

- Accélération des performances des grands modèles de langage environ 5,14 fois plus élevée avec les serveurs PowerEdge XE9680 par rapport à la génération précédente.
- Performances plus de dix fois supérieures attendues à l'avenir grâce aux optimisations du moteur NVIDIA Transformer Engine.
- Affinage ultrarapide des modèles d'apprentissage automatique à partir des jeux de données les plus récents.
- Gain d'espace dans le datacenter et refroidissement efficace avec un format 6U, par rapport au format 8U standard.

Aperçu des solutions

- Serveurs Dell PowerEdge XE9680 équipés de processeurs graphiques NVIDIA® H100
- Dell ProSupport

DELL Technologies

CyberAgent, Inc. est une société leader sur le marché japonais de la publicité sur Internet. Elle exploite plusieurs plateformes et services, y compris la plateforme de streaming innovante ABEMA. En 2016, la société a fondé AI Lab, un centre de recherche sur l'IA. Depuis, elle se consacre activement à la recherche et au développement de l'IA. En 2020, CyberAgent a mis au point une IA prédictive de pointe qui améliore la production de combinaisons percutantes de slogans publicitaires et d'images, boostant ainsi l'efficacité des publicités.

CyberAgent a poursuivi son développement de l'IA générative en créant un grand modèle de langage (LLM) unique, spécifique à la langue japonaise, comportant 13 milliards de paramètres. Conçu comme un modèle d'IA polyvalent pouvant être utilisé dans diverses situations, ce LLM peut être affiné pour créer des phrases d'accroche qui trouvent un écho auprès des utilisateurs de chaque plateforme publicitaire. CyberAgent utilise déjà son LLM japonais dans des services d'IA comme Kiwami Prediction AI, Kiwami Prediction TD et Kiwami Prediction LP pour assister la production de publicités créatives et prédire leur efficacité. À l'avenir, CyberAgent cherche à développer une IA multimodale capable de gérer non seulement les LLM japonais, mais aussi des images.

En mai 2023, CyberAgent a présenté un LLM japonais Open Source commercialisé sous le nom d'OpenCALM (Open CyberAgent Language Models), comportant 6,8 milliards de paramètres.

> **« Nos chercheurs internes peuvent sécuriser un plus grand nombre de ressources et les utiliser sans se soucier du coût, alors qu'auparavant, ils étaient incapables de sécuriser les processeurs graphiques dans le Cloud public ou devaient payer davantage pour une utilisation à long terme. »**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

Si ChatGPT est optimisé pour le chat, l'outil OpenCALM, lui, est plutôt un modèle de langage japonais générique qui peut être affiné pour répondre aux besoins des utilisateurs. CyberAgent a publié OpenCALM sous forme de projet Open Source, car il est plus avantageux pour l'entreprise de recevoir des retours d'informations d'autres sources et de collaborer avec d'autres sociétés pour contribuer au développement de la technologie de l'IA au Japon, plutôt que de développer un LLM japonais dans un environnement fermé.

L'infrastructure qui favorise l'innovation de CyberAgent en matière d'IA

Lorsque CyberAgent a créé son laboratoire AI Lab en 2016, chaque chercheur disposait pour ses recherches d'une station de travail équipée de processeurs graphiques. Or, en raison du télétravail imposé pendant la pandémie de 2020, il est devenu difficile de mettre à la disposition de chaque chercheur une station de travail dotée de processeurs graphiques. Pour veiller à ce que les chercheurs aient à leur disposition les ressources informatiques nécessaires, la société avait commencé à envisager la création de plateformes d'apprentissage automatique (ML) centralisées, reposant sur des serveurs optimisés par des processeurs graphiques, soit dans ses datacenters, soit dans le Cloud public. C'est alors que les tout derniers processeurs graphiques NVIDIA® A100 ont été commercialisés.

Daisuke Takahashi, Solution Architect, CIU, Group IT Department chez CyberAgent, Inc. explique : « Nous aurions pu opter pour un Cloud public si notre seul objectif avait été l'utilisation de processeurs graphiques, mais avec un Cloud public, on ne peut jamais savoir quand sera disponible la toute dernière génération de processeurs graphiques. De plus, nous n'avions aucune garantie que les processeurs graphiques seraient disponibles quand nous en aurions besoin. C'est pourquoi nous avons décidé de déployer des ressources de processeurs graphiques sur site, faciles à utiliser. Pour tirer parti de la flexibilité de l'infrastructure permettant de jongler entre Clouds public et privé, nous avons conçu une interface utilisateur aussi proche que possible des spécifications du Cloud public. » CyberAgent a construit sa plateforme ML initiale sur site en utilisant des serveurs Dell PowerEdge XE8545 équipés de quatre processeurs graphiques NVIDIA A100.

Pourquoi le choix de CyberAgent s'est porté sur des serveurs PowerEdge XE9680 équipés de processeurs graphiques NVIDIA H100

CyberAgent a continué à suivre les dernières innovations en matière de processeurs graphiques, notamment la sortie du tout dernier modèle NVIDIA H100. « Nous l'avons trouvé intéressant, pas uniquement pour ses performances améliorées, mais également pour ses mécanismes qui accélèrent des algorithmes de calcul spécifiques, avec par exemple le moteur Transformer Engine », ajoute M. Takahashi. « D'après NVIDIA, le moteur Transformer Engine permet un entraînement IA des LLM jusqu'à neuf fois plus rapide et des performances d'inférence de l'IA jusqu'à 30 fois plus élevées rapport aux processeurs graphiques NVIDIA A100 de la génération précédente. »

CyberAgent a opté pour le modèle de serveur PowerEdge XE9680 équipé de huit processeurs graphiques NVIDIA H100. M. Takahashi poursuit : « Lorsque nous avons eu connaissance du lancement de serveurs Dell PowerEdge XE9680 dotés de processeurs graphiques NVIDIA H100, nous avons décidé d'adopter cette solution sans tarder. Nous avons pu travailler en étroite collaboration avec Dell Technologies pour déterminer quelles configurations nous pouvions envisager avec les serveurs PowerEdge XE9680 et les processeurs graphiques qui allaient être commercialisés. Nous cherchions à augmenter le temps d'activité avec le moins d'unités possible. Nous nous sommes donc félicités en constatant que Dell Technologies pouvait nous fournir un niveau élevé de services de maintenance, y compris un service sur site dans les quatre heures, à un prix raisonnable. »

DELL Technologies

Accélération des performances d'un LLM comportant 13 milliards de paramètres 5,14 fois plus élevée aujourd'hui, et jusqu'à dix fois plus élevée à l'avenir.

M. Takahashi continue : « Si notre choix s'est porté sur les serveurs PowerEdge XE9680, c'est également parce que nos serveurs PowerEdge XE8545 existants offraient des performances stables et une facilité de maintenance. Par ailleurs, nous apprécions la facilité d'utilisation de l'outil de gestion Dell iDRAC qui permet une gestion sécurisée des serveurs en local et à distance. »

M. Takahashi a également apprécié la rapidité de déploiement : avec une commande passée en mars 2023, le déploiement s'est achevé un peu plus d'un mois plus tard, à la mi-mai. « Les chaînes logistiques étant sérieusement désorganisées en raison de la pandémie, j'ai également été rassuré par le fait que la chaîne logistique de Dell Technologies soit restée relativement stable. Il était bon de savoir que le déploiement pourrait se faire dans un délai si court. »

Plusieurs innovations ont été introduites dans le processus de mise en œuvre après le déploiement. M. Takashi se souvient : « Pour un LLM comportant un grand nombre de paramètres, nous avions besoin d'utiliser plusieurs processeurs graphiques. Nous avons donc installé huit cartes NIC de 400 Gbit/s dans chaque serveur et utilisé la technologie RDMA (Remote Direct Memory Access) pour créer une interconnexion haut débit entre les serveurs. Comme les serveurs équipés de processeurs graphiques génèrent beaucoup de chaleur, il est important d'intégrer dans leur conception un refroidissement efficace. Le format 6U des serveurs PowerEdge XE9680 offre un refroidissement robuste, ce qui est également appréciable. De plus, le datacenter a également été relocalisé vers un nouvel emplacement : des échangeurs thermiques y sont disponibles sur la porte arrière pour un refroidissement efficace reposant sur un système de refroidissement à l'eau à l'arrière des racks, plutôt que sur un système de refroidissement de toute la pièce abritant notre datacenter. »

Précision des slogans accrue grâce aux optimisations du moteur Transformer Engine

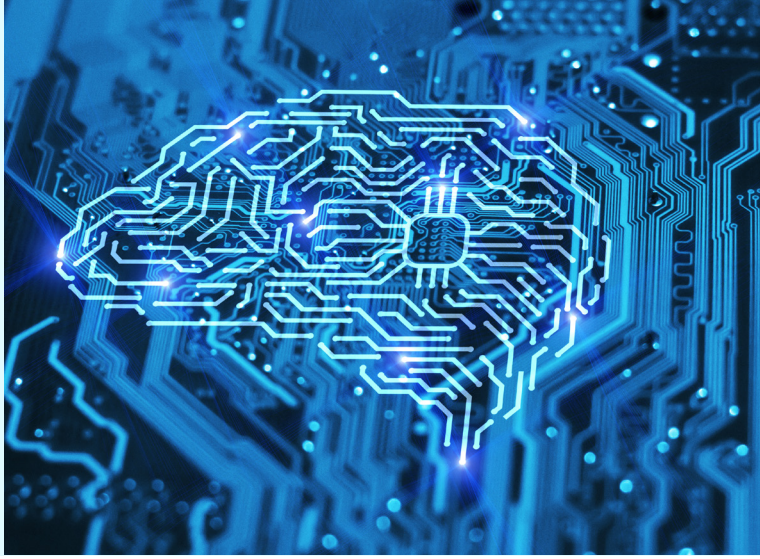
L'installation de serveurs PowerEdge XE9680 procure de nombreux avantages à CyberAgent. « Nous espérons pouvoir mettre à jour nos LLM japonais plus rapidement et plus fréquemment grâce à l'amélioration considérable des performances », indique M. Takahashi. « La vitesse de l'évolution des LLM japonais va également croître. En outre, par rapport aux serveurs PowerEdge XE8545 équipés de quatre processeurs graphiques NVIDIA A100, les serveurs PowerEdge XE9680 dotés de huit processeurs graphiques NVIDIA H100 offrent une amélioration des performances environ 5,14 fois supérieure. Nous prévoyons également des performances plus de dix fois supérieures à l'avenir grâce à l'optimisation du moteur NVIDIA Transformer Engine. Nous pouvons également affiner extrêmement rapidement les modèles de ML à partir des jeux de données les plus récents, ce qui nous permettra de répondre plus facilement aux demandes pour faire évoluer nos services, améliorer la précision des slogans et renforcer l'efficacité des contenus ».

L'infrastructure ML optimisée par des serveurs PowerEdge XE9680 a fait l'objet de vifs éloges de la part des utilisateurs. « Nos chercheurs internes nous ont rapporté qu'ils pouvaient sécuriser un plus grand nombre de ressources et les utiliser sans se soucier du coût, alors qu'auparavant, ils étaient incapables de sécuriser les processeurs graphiques dans le Cloud public ou devaient payer davantage pour une utilisation à long terme », déclare M. Takahashi. « Autre avantage : nous avons pu fournir une infrastructure répondant aux spécifications les plus strictes, y compris en matière d'interconnexion, afin que les utilisateurs puissent générer des résultats pour l'entreprise. »

M. Takahashi apprécie également l'outil de gestion iDRAC de Dell Technologies, que l'entreprise utilise depuis un certain temps, car il réduit la charge de gestion. « Nous ne sommes pas toujours sur le site du datacenter, c'est pourquoi l'outil iDRAC est utile pour effectuer des tâches à distance, par exemple vérifier la température et l'état des processeurs graphiques et mettre à jour le firmware sans avoir à accéder au système d'exploitation. »

> **« Le format 6U des serveurs PowerEdge XE9680 offre un refroidissement robuste, ce qui est également appréciable. »**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

> **Nous espérons pouvoir mettre à jour nos LLM japonais plus rapidement. Les serveurs PowerEdge XE9680 équipés de huit processeurs NVIDIA H100 ont permis des performances environ 5,14 fois supérieures. »**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.



Efforts axés sur les LLM, les processeurs graphiques et l'infrastructure

À l'avenir, CyberAgent prévoit d'exploiter les commentaires reçus et les données d'apprentissage collectées à partir d'OpenCALM pour améliorer le LLM qu'utilisent ses collaborateurs. Via OpenCALM, CyberAgent explore également les possibilités de collaboration avec des sociétés et organisations dans des secteurs d'activité autres que la publicité. Par exemple, CyberAgent a engagé un dialogue avec des acteurs du commerce de détail et de la finance pour créer des LLM sectoriels qui apprennent des données spécifiques à un secteur d'activité.

En attendant, M. Takahashi explique qu'il restera à l'affût pour se tenir au courant de la commercialisation des processeurs graphiques les plus récents et des nouvelles technologies connexes. « Nous sommes également impatients de voir comment d'autres fournisseurs parviendront à créer un écosystème logiciel similaire à celui de NVIDIA. Je suis également intéressé par l'implémentation de NVIDIA NVLink-C2C et de nouveaux standards comme CXL (Compute eXpress Link) qui connectent le processeur et le processeur graphique, car le bus PCIe peut constituer un goulot d'étranglement qui freine les performances des processeurs graphiques. J'espère que Dell Technologies continuera d'adopter de nouvelles technologies à un rythme rapide et de concevoir des produits qui tiennent leurs promesses en matière de performances. »

En utilisant les processeurs graphiques les plus récents et les plus rentables, l'équipe de recherche et développement en IA de CyberAgent continuera d'évoluer en fournissant une infrastructure ML répondant aux exigences des utilisateurs. En outre, avec les prochaines évolutions du LLM japonais, CyberAgent continuera à susciter un fort intérêt, non seulement dans sa propre activité de services de publicité, mais aussi sur le marché japonais de l'IA.

> **Nous cherchions à augmenter le temps d'activité avec le moins d'unités possible. Nous nous sommes donc félicités en constatant que Dell Technologies pouvait nous fournir un niveau élevé de services de maintenance, y compris un service sur site dans les quatre heures, à un prix raisonnable. »**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.

Ce contenu a été traduit à partir de la version japonaise par Dell Technologies

En savoir plus sur les solutions d'IA générative Dell Technologies.

Suivez-nous sur les réseaux sociaux.

DELL Technologies