DELL Technologies

MEDIA E INTRATTENIMENTO
GIAPPONE

CyberAgent

# Sviluppo dell'AI generativa giapponese e trasformazione dei servizi pubblicitari digitali

CyberAgent, Inc. utilizza i server Dell PowerEdge XE9680 con otto GPU NVIDIA® H100 Tensor Core per accelerare l'AI generativa e migliorare l'efficacia della pubblicità.

## Esigenze aziendali

Dal 2016, CyberAgent, Inc. è attiva nella ricerca e nello sviluppo dell'AI e integra questa tecnologia nella sua attività pubblicitaria. L'azienda doveva fornire al personale accesso rapido ed economico a server on-premise altamente affidabili con le GPU NVIDIA più avanzate in commercio per le sue iniziative di sviluppo di AI generativa.

## Risultati di business

- Accelera le prestazioni dei modelli linguistici di grandi dimensioni (LLM) di circa 5,14 volte rispetto alla generazione precedente con server PowerEdge XE9680.
- Prevede un miglioramento futuro delle prestazioni di oltre 10 volte con le ottimizzazioni NVIDIA Transformer Engine.
- Consente l'ottimizzazione ad alta velocità dei modelli di apprendimento automatico secondo i data set più recenti.
- Risparmia spazio nel data center e ottiene un raffreddamento efficiente con un fattore di forma 6U rispetto al classico 8U.

## Soluzioni in breve

- Server Dell PowerEdge XE9680 con GPU NVIDIA® H100
- Dell ProSupport

DELL Technologies

CyberAgent, Inc. è un'azienda conosciuta come leader di mercato nel settore nazionale della pubblicità su Internet e per iniziative come l'innovativa piattaforma TV, ABEMA. Nel 2016, l'azienda ha istituito un'organizzazione di ricerca AI chiamata AI Lab e da allora è attiva nella ricerca e nello sviluppo dell'AI. Nel 2020, CyberAgent ha introdotto un'AI predittiva all'avanguardia che migliora la produzione di slogan per banner pubblicitari e combinazioni di immagini ad alto impatto, incrementando l'efficacia della pubblicità.

CyberAgent porta avanti lo sviluppo dell'AI generativa, creando un esclusivo modello linguistico di grandi dimensioni (LLM) giapponese con 13 miliardi di parametri. Concepito come modello di AI per scopi generici da poter utilizzare in una varietà di situazioni, l'LLM può essere ottimizzato per creare testi accattivanti in linea con gli utenti di ciascuna piattaforma pubblicitaria. CyberAgent utilizza già il suo LLM giapponese in servizi di AI come Kiwami Prediction AI, Kiwami Prediction TD e Kiwami Prediction LP per supportare la produzione creativo-pubblicitaria e prevedere l'efficacia della pubblicità. In futuro, CyberAgent mira a sviluppare un'AI multimodale che possa gestire non solo gli LLM giapponesi ma anche le immagini.

> **"I nostri ricercatori interni possono proteggere una maggiore quantità di risorse e utilizzarle senza preoccuparsi dei costi, mentre prima non potevano proteggere le GPU nel public cloud o pagavano di più per un utilizzo a lungo termine."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

A maggio 2023, CyberAgent ha rilasciato un LLM giapponese open source disponibile in commercio denominato OpenCALM (Open CyberAgent Language Models), che include fino a 6,8 miliardi di parametri.

Mentre la ChatGPT è ottimizzata per le conversazioni, OpenCALM è un modello linguistico giapponese con finalità più generiche che può essere ottimizzato per rispondere alle esigenze degli utenti. CyberAgent ha rilasciato OpenCALM come progetto open source poiché per l'azienda è più vantaggioso ricevere il feedback da altre fonti e collaborare con altre aziende per contribuire allo sviluppo della tecnologia AI in Giappone, piuttosto che sviluppare un LLM giapponese in un ambiente chiuso.

## L'infrastruttura alla base dell'innovazione AI di CyberAgent

Quando CyberAgent ha istituito AI Lab nel 2016, ogni ricercatore aveva una workstation basata su GPU per condurre le ricerche. Tuttavia, la necessità di lavorare da remoto durante la pandemia del 2020 ha reso difficile per ogni ricercatore utilizzare workstation basate su GPU. Per garantire che i ricercatori avessero le risorse aziendali necessarie, l'azienda ha iniziato a pensare alla creazione di piattaforme di apprendimento automatico (ML) centralizzate con server basati su GPU nei suoi data center o nel public cloud quando le più recenti GPU NVIDIA® A100 sono state rilasciate.

Daisuke Takahashi, Solution Architect, CIU, Group IT Department di CyberAgent, Inc. afferma: "Avremmo potuto scegliere un public cloud se avessimo voluto solo utilizzare GPU, ma con un public cloud non si sa mai quando verranno rilasciate nuove GPU. Inoltre, non era garantito che le GPU sarebbero state disponibili al momento necessario, quindi abbiamo deciso di implementare risorse GPU on-premise con facilità d'uso. Per ottenere la flessibilità dell'infrastruttura per passare dal public cloud al private cloud e viceversa, abbiamo concepito un'interfaccia utente quanto più possibile vicina alle specifiche del public cloud". CyberAgent ha creato la sua piattaforma ML on-premise iniziale utilizzando server Dell PowerEdge XE8545 con quattro GPU NVIDIA A100.

## Perché CyberAgent ha scelto i server PowerEdge XE9680 con GPU NVIDIA H100

CyberAgent ha continuato a seguire l'innovazione delle GPU, soprattutto le ultime GPU NVIDIA H100. "Erano interessanti non solo per le prestazioni migliorate, ma anche per meccanismi come Transformer Engine che accelerano specifici algoritmi computazionali", spiega Takahashi. "Secondo NVIDIA, Transformer Engine può accelerare l'addestramento AI degli LLM fino a nove volte e l'inferenza AI fino a 30 volte rispetto alle GPU NVIDIA A100 della generazione precedente."

CyberAgent ha scelto il modello di server PowerEdge XE9680 con otto GPU NVIDIA H100. Takahashi spiega: "Quando abbiamo saputo del rilascio dei server Dell PowerEdge XE9680 con GPU NVIDIA H100, abbiamo deciso di adottarli il prima possibile. Ci siamo confrontati con Dell Technologies in merito alle configurazioni possibili con i server PowerEdge XE9680 e le GPU in arrivo. Volevamo aumentare l'uptime con il minor numero possibile di unità, pertanto siamo stati soddisfatti che Dell Technologies ci abbia fornito un elevato livello di manutenzione, tra cui il servizio on-site entro quattro ore, a un prezzo ragionevole".

DELL Technologies

**Accelera un LLM con 13 miliardi di parametri di 5,14 volte oggi e più di 10 volte in futuro.**

Takahashi continua: "Abbiamo scelto i server PowerEdge XE9680 anche perché le precedenti installazioni di server PowerEdge XE8545 hanno garantito prestazioni stabili e facilità di manutenzione. Inoltre, diamo importanza alla facilità d'uso dello strumento di gestione Dell iDRAC per la gestione dei server locale e remota".

Takahashi apprezza il fatto che, a seguito l'ordine effettuato a marzo 2023, la consegna è stata completata poco più di un mese dopo, a metà maggio. "Con la confusione che la pandemia ha causato nelle supply chain, sono stato inoltre rassicurato dal fatto che Dell Technologies avesse una supply chain relativamente stabile ed è stato positivo sapere che avrebbe potuto consegnare in così poco tempo."

Una varietà di innovazioni è stata apportata al processo di creazione post-consegna. Takahashi ricorda: "Per un LLM con un elevato numero di parametri, dovevamo usare diverse GPU, quindi abbiamo installato otto schede di interfaccia di rete (NIC) da 400 Gbps su ciascun server e abbiamo utilizzato la tecnologia RDMA (Remote Direct Memory Access) per creare un'interconnessione ad alta velocità tra i server. I server GPU generano molto calore per cui è importante che siano concepiti per essere raffreddati in modo efficiente. Il fattore di forma 6U dei server PowerEdge XE9680 per il raffreddamento solido è un altro elemento eccellente. Inoltre, il data center è stato ricollocato in una nuova posizione dove ci sono scambiatori di calore con porta posteriore. Abbiamo ottenuto un raffreddamento efficace attraverso l'installazione di scambiatori di calore con porta posteriore e raffreddamento ad acqua sul retro dei rack, anziché raffreddare l'intera stanza che ospita il data center".

## Migliore accuratezza degli slogan con le ottimizzazioni Transformer Engine

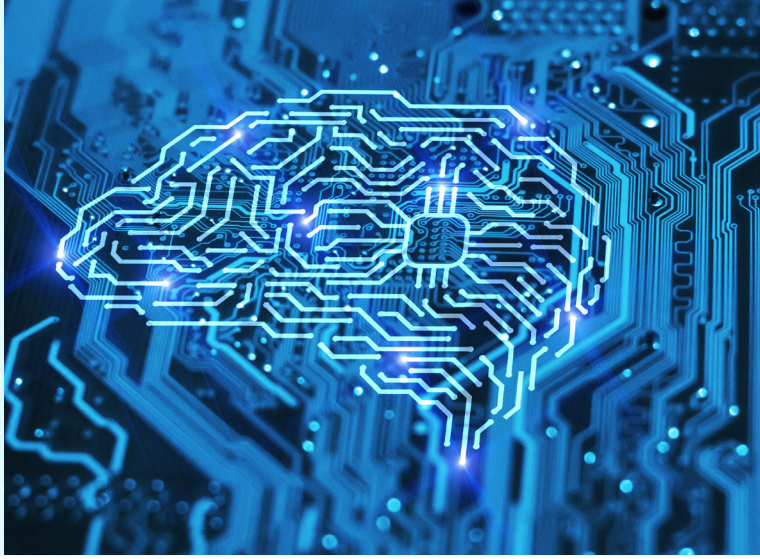
Con l'installazione dei server PowerEdge XE9680, CyberAgent ottiene una serie di vantaggi. "Ci aspettiamo di poter aggiornare i nostri LLM giapponesi più velocemente e più spesso grazie a un importante miglioramento delle prestazioni", dichiara Takahashi. "Migliorerà anche la velocità dell'evoluzione degli LLM giapponesi. Inoltre, rispetto ai server PowerEdge XE8545 con quattro GPU NVIDIA A100, i server PowerEdge XE9680 con otto GPU NVIDIA H100 hanno ottenuto un miglioramento delle prestazioni di circa 5,14 volte. Prevediamo anche un aumento delle prestazioni di oltre 10 volte attraverso l'ottimizzazione di NVIDIA Transformer Engine in futuro. Possiamo inoltre eseguire l'ottimizzazione ad alta velocità dei modelli ML secondo i più recenti data set, il che renderà più semplice rispondere alle richieste di evoluzione dei nostri servizi, migliorare l'accuratezza degli slogan e distribuire contenuti più efficaci."

L'infrastruttura ML basata su server PowerEdge XE9680 è stata molto apprezzata dagli utenti. "I nostri ricercatori interni riferiscono di poter proteggere una maggiore quantità di risorse e utilizzarle senza preoccuparsi dei costi, mentre prima non potevano proteggere le GPU nel public cloud o pagavano di più per un utilizzo a lungo termine", dichiara Takahashi. "Un altro vantaggio è la possibilità di fornire un'infrastruttura con specifiche elevate, compresa l'interconnessione, in modo che gli utenti possano avere un impatto aziendale."

Takahashi apprezza anche lo strumento di gestione Dell Technologies iDRAC, che l'azienda utilizza da tempo, poiché riduce l'onere di gestione. "Non essendo sempre presenti nel data center, iDRAC è utile per le operazioni da remoto, ad esempio controllare la temperatura e lo stato delle GPU e aggiornare il firmware senza dover accedere al sistema operativo."

> **"Il fattore di forma 6U dei server PowerEdge XE9680 per il raffreddamento solido è un altro elemento eccellente."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

> **Ci aspettiamo di poter aggiornare i nostri LLM giapponesi più velocemente. I server PowerEdge XE9680 con otto GPU NVIDIA H100 hanno raggiunto un miglioramento delle prestazioni di circa 5,14 volte."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.



## Attenzione rivolta a LLM, GPU e infrastruttura

In futuro, CyberAgent prevede di utilizzare il feedback e gli insegnamenti ottenuti da OpenCALM per migliorare l'LLM che utilizzano i suoi dipendenti. Attraverso OpenCALM, CyberAgent sta esplorando anche delle collaborazioni con aziende e organizzazioni in settori diversi da quello pubblicitario. Ad esempio, CyberAgent ha avviato discussioni con attori nelle aree retail e finanziaria per creare LLM specifici per il loro ambito che apprendano dai dati del settore.

Allo stesso tempo, Takahashi spiega che continuerà ad aggiornarsi con le più recenti GPU e le nuove tecnologie correlate per vedere come vengono commercializzate. "Attendiamo inoltre di vedere in che modo altri vendor possono creare un ecosistema software simile a quello raggiunto da NVIDIA. Sono interessato anche all'implementazione di NVIDIA NVLink-C2C e nuovi standard come CXL (Compute eXpress Link) che connettono CPU e GPU, poiché il bus PCIe potrebbe essere un collo di bottiglia per le prestazioni delle GPU. Credo che Dell Technologies continuerà ad adottare nuove tecnologie a un ritmo sostenuto e a progettare prodotti che assicurano prestazioni elevate."

Utilizzando le GPU più recenti e a costi contenuti, il team di ricerca e sviluppo AI di CyberAgent continuerà a evolvere fornendo l'infrastruttura ML che richiedono gli utenti. Inoltre, con l'ulteriore sviluppo dell'LLM giapponese, CyberAgent continuerà ad attirare un'attenzione significativa, non solo sulla sua attività pubblicitaria ma anche sul mercato dell'AI giapponese.

> **Volevamo aumentare l'uptime con il minor numero possibile di unità, pertanto siamo stati soddisfatti che Dell Technologies ci abbia fornito un elevato livello di manutenzione, tra cui il servizio on-site entro quattro ore, a un prezzo ragionevole".**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.

Questo contenuto è stato tradotto da Dell Technologies dal giapponese.

**Scopri di più** sulle soluzioni di AI generativa Dell Technologies.

**Seguici** sui social.

DELL Technologies

Copyright © 2023 Dell Inc. o sue società controllate. Tutti i diritti riservati. Dell Technologies, Dell e altri marchi sono marchi registrati di Dell Inc. o delle sue società controllate. Gli altri marchi appartengono ai rispettivi proprietari. Questo caso di studio ha scopo puramente informativo. Dell ritiene che le informazioni presenti in questo documento siano accurate alla data di pubblicazione (settembre 2023). Le informazioni sono soggette a modifiche senza preavviso. Dell non offre garanzie di alcun tipo, espresse o implicite, per questo caso di studio.