DELL Technologies

MEDIEN UND UNTERHALTUNG
JAPAN

CyberAgent

Entwicklung japanischer generativer KI und Transformation digitaler Werbeservices

CyberAgent Inc. kombiniert Dell PowerEdge XE9680-Server mit acht NVIDIA® H100 Tensor Core-GPUs, um generative KI zu beschleunigen und die Wirksamkeit von Werbung zu verbessern.

Geschäftsanforderungen

Seit 2016 forscht und entwickelt CyberAgent Inc. aktiv auf dem Gebiet der KI und bindet die Technologie in sein Werbegeschäft ein. Die MitarbeiterInnen des Unternehmens benötigten schnellen, kosteneffizienten Zugang zu hochzuverlässigen On-Premise-Servern, die die Entwicklung generativer KI mit den modernsten NVIDIA-GPUs unterstützen.

Geschäftsergebnisse

- Annähernd 5,14-fache Performancesteigerung beim LLM (Large Language Model) mit PowerEdge XE9680-Servern im Vergleich zur vorherigen Servergeneration
- Erwartete Performancesteigerung um mehr als das 10-Fache durch künftige Optimierungen der NVIDIA Transformer Engine
- Unterstützung von Highspeed-Finetuning für ML-Modelle (maschinelles Lernen) anhand der neuesten Datasets
- Geringerer Platzbedarf im Rechenzentrum und effiziente Kühlung mit einem 6-HE-Formfaktor im Vergleich zum herkömmlichen 8-HE-Formfaktor

Lösungen auf einen Blick

- Dell PowerEdge XE9680-Server mit NVIDIA® H100-GPUs
- Dell ProSupport

CyberAgent Inc. hat sich in der japanischen Internet-Werbebranche und in Ventures als Marktführer behauptet und unter anderem die innovative TV-Plattform ABEMA entwickelt. Im Jahr 2016 gründete das Unternehmen die KI-Forschungseinrichtung „AI Lab“ und forscht und entwickelt seither aktiv im Bereich der KI. Im Jahr 2020 führte CyberAgent eine hochmoderne vorausschauende KI ein, die die Generierung von Schlagwort- und Bildkombinationen für aufmerksamkeitsstarke Bannerwerbung verbessert und so die Werbewirksamkeit erhöht.

CyberAgent arbeitete weiter im Bereich der generativen KI-Entwicklung und realisierte ein einzigartiges, auf die japanische Sprache zugeschnittenes Large Language Model (LLM) mit 13 Milliarden Parametern. Das LLM wurde als universelles KI-Modell für verschiedenste Zwecke entwickelt. Durch Finetuning generiert es Schlagwörter, mit denen NutzerInnen verschiedener Werbeplattformen gezielt angesprochen werden. CyberAgent nutzt sein japanisches LLM bereits in KI-Services wie Kiwami Prediction AI, Kiwami Prediction TD und Kiwami Prediction LP, um die Produktion kreativer Werbung zu unterstützen und die Werbewirksamkeit vorherzusagen. Für die Zukunft strebt CyberAgent die Entwicklung einer multimodalen KI an, die nicht nur japanische LLMs unterstützen, sondern auch Bilder verarbeiten kann.

> **Unser internes Forschungsteam kann größere Mengen an Ressourcen reservieren und ohne Kostenbedenken nutzen. In der Public Cloud war eine Reservierung von GPUs früher nicht möglich oder es wurden Mehrkosten für die langfristige Nutzung berechnet.“**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent Inc.

Im Mai 2023 brachte CyberAgent ein kommerzielles japanisches Open-Source-LLM namens OpenCALM (Open CyberAgent Language Models) heraus, das bis zu 6,8 Milliarden Parameter umfasst.

Während ChatGPT für Chats optimiert ist, dient OpenCALM als allgemeines japanisches Sprachmodell, das an die Anforderungen von NutzerInnen angepasst werden kann. CyberAgent hat OpenCALM als Open-Source-Projekt gestartet, um Feedback aus anderen Quellen zu erhalten und gemeinsam mit anderen Unternehmen zur Entwicklung der KI-Technologie in Japan beizutragen. Dies ist für das Unternehmen vorteilhafter als die Entwicklung eines japanischen LLM in einer geschlossenen Umgebung.

Die Infrastruktur hinter den KI-Innovationen von CyberAgent

Als CyberAgent 2016 sein AI Lab gründete, erhielt jedes Mitglied der Forschungsgruppe eine GPU-gestützte Workstation. Mit Beginn der Pandemie im Jahr 2020 und dem notwendigen Umstieg auf mobiles Arbeiten hatten die Mitglieder des Forschungsteams jedoch nur noch eingeschränkten Zugang zu ihren GPU-gestützten Workstations. Mit der Markteinführung der neuesten NVIDIA® A100-GPUs überlegte das Unternehmen, wie es seinem Forschungsteam die benötigten Computing-Ressourcen zur Verfügung stellen könnte. Der Ansatz war, zentralisierte ML-Plattformen (maschinelles Lernen) mit GPU-gestützten Servern in den eigenen Rechenzentren oder in der Public Cloud einzurichten.

Dazu Daisuke Takahashi, Solution Architect, CIU, Group IT Department bei CyberAgent Inc.: „Wenn es uns nur um GPUs gegangen wäre, hätten wir uns für eine Public Cloud entscheiden können. Allerdings man weiß nie, wann die neuesten GPUs Einzug in die Public Cloud halten. Ein weiterer Unsicherheitsfaktor war, ob die GPUs zum gewünschten Zeitpunkt verfügbar sein würden. Deshalb entschieden wir uns für die Bereitstellung leicht zugänglicher On-Premise-GPU-Ressourcen. Damit die Infrastruktur den flexiblen Wechsel zwischen Public Cloud und Private Cloud unterstützt, haben wir eine Benutzeroberfläche entwickelt, die den Spezifikationen der Public Cloud so nahe wie möglich kommt.“ CyberAgent realisierte seine erste On-Premise-ML-Plattform mit Dell PowerEdge XE8545-Servern, die mit vier NVIDIA A100-GPUs ausgestattet waren.

PowerEdge XE9680-Server mit NVIDIA H100-GPUs – die Entscheidungsgründe von CyberAgent

CyberAgent behielt die GPU-Innovationen weiterhin im Blick, insbesondere die neueste NVIDIA H100-GPU. „Die Lösung überzeugte nicht nur durch ihre verbesserte Performance, sondern auch durch Mechanismen wie die Transformer Engine, die bestimmte Rechenalgorithmen beschleunigt“, erklärt Takahashi. „Laut NVIDIA bietet die Transformer Engine gegenüber NVIDIA A100-GPUs der vorherigen Generation ein bis zu 9-mal schnelleres KI-Training von LLMs und eine bis zu 30-mal schnellere KI-Inferenz.“

CyberAgent entschied sich für das PowerEdge XE9680-Servermodell mit acht NVIDIA H100-GPUs. Takahashi ergänzt: „Als wir von der geplanten Markteinführung der Dell PowerEdge XE9680-Server mit NVIDIA H100-GPUs erfuhren, beschlossen wir, diese so schnell wie möglich einzusetzen. Wir standen in engem Austausch mit Dell Technologies, um die möglichen Konfigurationen der neuen PowerEdge XE9680-Server und GPUs zu besprechen. Unser Ziel war eine höhere Verfügbarkeit mit der Mindestanzahl an Geräten. Deshalb war das erstklassige Wartungsangebot von Dell Technologies für uns interessant, das einen Vor-Ort-Service innerhalb von vier Stunden zu einem angemessenen Preis beinhaltete.“

Derzeit 5,14-fache Beschleunigung der LLM-Performance bei 13 Milliarden Parametern, mehr als 10-fache Leistung prognostiziert.

Takahashi weiter: „Für die PowerEdge XE9680-Server sprachen außerdem die stabile Performance und die einfache Wartung, die wir von unseren installierten PowerEdge XE8545-Servern kannten. Darüber hinaus schätzen wir das benutzerfreundliche Dell iDRAC-Managementtool, das für die sichere Verwaltung von lokalen und Remoteservern sorgt."

Takahashi betont außerdem, dass die Bestellung im März 2023 aufgegeben wurde und Mitte Mai, also etwas mehr als einen Monat später, abgewickelt war. „Angesichts der pandemiebedingt unterbrochenen Lieferketten war ich auch erleichtert, dass Dell Technologies über eine relativ stabile Lieferkette verfügt. Es war beruhigend zu wissen, dass sie in so kurzer Zeit liefern können."

Nach der Auslieferung wurden beim Aufbau einige Innovationen implementiert. Takahashi blickt zurück: „Für ein LLM mit einer großen Anzahl von Parametern benötigten wir mehrere GPUs. Folglich installierten wir acht 400-Gbit/s-Netzwerkschnittstellenkarten (NICs) pro Server und stellten über die RDMA-Technologie (Remote Direct Memory Access) eine Highspeed-Interconnect-Verbindung zwischen den Servern her. GPU-Server erzeugen viel Wärme, sodass eine effiziente Kühlung äußerst wichtig ist. Auch hier punkten die PowerEdge XE9680-Server mit 6-HE-Formfaktor durch ihre zuverlässige Kühlung. Außerdem wurde das Rechenzentrum an einen neuen Standort verlegt, wo Rücktür-Wärmetauscher zur Verfügung stehen. Durch die Montage von wassergekühlten Rücktür-Wärmetauschern an der Rückseite der Racks erreichen wir eine effektive Kühlung, ohne den gesamten umgebenden Raum unseres Rechenzentrums zu kühlen."

Transformer Engine-Optimierungen für präzisere Schlagwörter

Die Installation von PowerEdge XE9680-Servern bietet CyberAgent mehrere Vorteile. „Wir gehen davon aus, dass wir unsere japanischen LLMs aufgrund der deutlichen Performancesteigerungen schneller und häufiger aktualisieren können", so Takahashi. „Bei der Entwicklung der japanischen LLMs werden wir auch an Tempo zulegen. Im Vergleich zu den PowerEdge XE8545-Servern mit vier NVIDIA A100-GPUs erzielten die PowerEdge XE9680-Server mit acht NVIDIA H100-GPUs eine etwa 5,14-fache Performancesteigerung. Durch künftige Optimierungen der NVIDIA Transformer Engine rechnen wir außerdem mit einem mehr als 10-fachen Performancezuwachs. Zudem können wir anhand der neuesten Datasets ein Highspeed-Finetuning für ML-Modelle durchführen und so die Nachfrage nach verbesserten Services leichter bedienen, die Genauigkeit von Schlagwörtern verbessern und effektivere Inhalte bereitstellen."

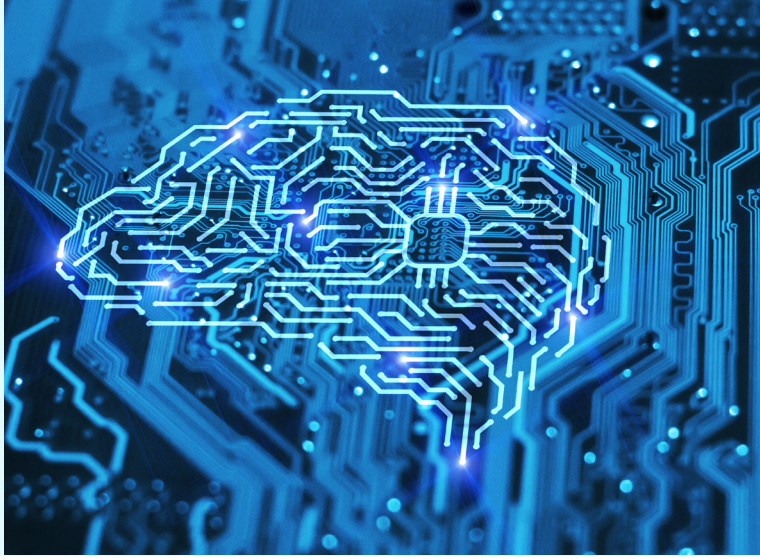
Die von PowerEdge XE9680-Servern unterstützte ML-Infrastruktur wird auch von Nutzerseite ausgesprochen positiv bewertet. „Unser internes Forschungsteam berichtet, dass es größere Mengen an Ressourcen reservieren und ohne Kostenbedenken nutzen kann. In der Public Cloud war eine Reservierung von GPUs früher nicht möglich oder es wurden Mehrkosten für die langfristige Nutzung berechnet", sagt Takahashi. „Ein weiterer Vorteil ist die hoch spezialisierte Infrastruktur einschließlich Interconnect-Technologie, durch die die NutzerInnen ihren Geschäftserfolg steigern können."

Takahashi schätzt außerdem das iDRAC-Verwaltungstool von Dell Technologies, das im Unternehmen schon seit einiger Zeit für weniger Verwaltungsaufwand sorgt. „Wir sind nicht immer vor Ort im Rechenzentrum, sodass sich die Remotefunktionen von iDRAC bereits bewährt haben, wie etwa die Temperatur- und Zustandskontrolle der GPUs sowie Firmwareupdates, ohne auf das Betriebssystem zugreifen zu müssen."

> **„Die PowerEdge XE9680-Server mit 6-HE-Formfaktor zeichnen sich zudem durch ihre zuverlässige Kühlung aus."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent Inc.

> **Wir gehen davon aus, dass wir unsere japanischen LLMs schneller aktualisieren können. Die PowerEdge XE9680-Server mit acht NVIDIA H100-GPUs erzielten eine annähernd 5,14-fache Performancesteigerung.“**

Daisuke Takahashi
Solution Architect, CIU, Group IT Department,
CyberAgent Inc.



LLMs, GPUs und Infrastruktur im Fokus

Für die Zukunft plant CyberAgent, das Feedback und die Erkenntnisse aus OpenCALM zu nutzen, um das LLM für seine MitarbeiterInnen zu verbessern. Über OpenCALM evaluiert CyberAgent auch die mögliche Zusammenarbeit mit Unternehmen und Organisationen aus anderen Bereichen als der Werbebranche. So führt CyberAgent bereits Gespräche mit dem Einzelhandel und mit Finanzdienstleistern, um branchenspezifische LLMs zu entwickeln, die wiederum aus branchenspezifischen Daten lernen.

In der Zwischenzeit, so Takahashi, beobachtet sein Unternehmen weiterhin den Markt für die neuesten GPUs und die damit verbundenen innovativen Technologien. „Wir sind auch gespannt auf die Erfolge anderer Anbieter, die ein ähnliches Softwareökosystem wie das von NVIDIA aufbauen. Ein interessantes Thema ist zudem die Implementierung von NVIDIA NVLink-C2C und neuen Standards wie CXL (Compute eXpress Link) zur CPU-GPU-Verbindung, da der PCIe-Bus die GPU-Performance beeinträchtigen kann. Ich erwarte, dass Dell Technologies auch künftig neue Technologien in raschem Tempo einführen und leistungsfähige Produkte entwickeln wird.“

Durch den Einsatz modernster, kosteneffektiver GPUs beabsichtigt das KI-Forschungs- und Entwicklungsteam von CyberAgent, die ML-Infrastruktur im Einklang mit neuen Nutzeranforderungen weiter auszubauen. Darüber hinaus wird CyberAgent durch die Weiterentwicklung des japanischen LLM weiterhin im Blickpunkt stehen, nicht nur im Werbegeschäft, sondern auch auf dem japanischen KI-Markt.

> **Unser Ziel war eine höhere Verfügbarkeit mit der Mindestanzahl an Geräten. Deshalb war das erstklassige Wartungsangebot von Dell Technologies für uns interessant, das einen Vor-Ort-Service innerhalb von vier Stunden zu einem angemessenen Preis beinhaltete.“**

Daisuke Takahashi
Solution Architect, CIU, Group IT Department,
CyberAgent Inc.

Dieser Inhalt wurde von Dell Technologies aus dem Japanischen übersetzt.

Weitere Informationen zu Dell Technologies Generative AI Solutions.

Auf Social Media **folgen**

DELL Technologies