DELL Technologies

MEDIA AND ENTERTAINMENT
JAPAN

CyberAgent

De Japanse Generative AI ontwikkelen en digitale reclamediensten transformeren

CyberAgent, Inc. gebruikt Dell PowerEdge XE9680 servers met acht NVIDIA® H100 Tensor Core GPU's om Generative AI te versnellen en de doeltreffendheid van advertenties te verbeteren.

Bedrijfsbehoeften

Sinds 2016 is CyberAgent, Inc. actief bezig met het onderzoeken, ontwikkelen en integreren van AI in zijn reclameactiviteiten. Het bedrijf moest zijn personeel snelle en betaalbare toegang bieden tot zeer betrouwbare servers on-premises met de meest geavanceerde NVIDIA GPU's die beschikbaar zijn voor het ontwikkelen van Generative AI.

Bedrijfsresultaten

- Versnelt de prestaties van het grote taalmodel (LLM) met ongeveer 5,14 keer vergeleken met de vorige generatie met PowerEdge XE9680 servers.
- Verwacht een prestatieverbetering van meer dan 10 keer in de toekomst met NVIDIA Transformer Engine optimalisaties.
- Verfijnt op hoge snelheid modellen voor machine learning op basis van de nieuwste datasets.
- Bespaart ruimte in het datacenter en levert efficiënte koeling met een vormfactor van 6U in plaats van de gebruikelijke 8U.

Overzicht van oplossingen

- Dell PowerEdge XE9680 servers met NVIDIA® H100 GPU's
- Dell ProSupport

CyberAgent, Inc. is een bedrijf dat bekend staat als marktleider in de binnenlandse industrie en ondernemingen van internetreclame, waaronder het innovatieve tv-platform ABEMA. In 2016 richtte het bedrijf een AI-onderzoeksorganisatie op met de naam AI Lab en sindsdien is het actief bezig met onderzoek naar en de ontwikkeling van AI. In 2020 introduceerde CyberAgent een geavanceerde, voorspellende AI die de productie van opvallende slogans en beeldcombinaties voor banneradvertenties verbetert, waardoor de advertenties veel effectiever worden.

CyberAgent heeft de ontwikkeling van haar Generative AI voortgezet door een uniek groot taalmodel (LLM) voor de Japanse taal met 13 miljard parameters te maken. Deze LLM is ontworpen als een AI-model voor algemeen gebruik dat in verschillende situaties kan worden toegepast en kan worden verfijnd om slogans te creëren die aanslaan bij de gebruikers van elk advertentieplatform. CyberAgent gebruikt haar Japanse LLM in AI-diensten zoals Kiwami Prediction AI, Kiwami Prediction TD en Kiwami Prediction LP om de productie van creatieve reclame te ondersteunen en de effectiviteit van reclames te voorspellen. In de toekomst wil CyberAgent een multimodale AI ontwikkelen die niet alleen overweg kan met Japanse LLM's, maar ook met afbeeldingen.

> **"Onze interne onderzoekers kunnen een grotere hoeveelheid bronnen veiligstellen en gebruiken zonder zich zorgen te hoeven maken over de kosten, terwijl ze voorheen geen GPU's konden veiligstellen in de public cloud of meer moesten betalen voor langdurig gebruik."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.

In mei 2023 heeft CyberAgent een commercieel beschikbare open-source Japanse LLM genaamd OpenCALM (Open CyberAgent Language Models) uitgebracht die tot 6,8 miljard parameters bevat.

Terwijl ChatGPT is bedoeld voor chatten, is OpenCALM meer een algemeen Japans taalmodel dat kan worden afgestemd op de behoeften van gebruikers. CyberAgent heeft OpenCALM uitgebracht als een open-source-project omdat het voor het bedrijf gunstiger is feedback te ontvangen van andere bronnen en samen te werken met andere bedrijven om bij te dragen aan de ontwikkeling van de AI-technologie in Japan, dan om een Japanse LLM in een gesloten omgeving te ontwikkelen.

De infrastructuur die de AI-innovatie van CyberAgent mogelijk maakt

Toen CyberAgent in 2016 zijn AI-lab oprichtte, had iedere onderzoeker een door GPU aangestuurd werkstation voor onderzoek. De noodzaak om tijdens de pandemie van 2020 op afstand te gaan werken, maakte het echter voor iedere onderzoeker moeilijk om hun werk met door GPU aangedreven werkstations te kunnen doen. Om ervoor te zorgen dat onderzoekers konden beschikken over de rekenkracht die ze nodig hadden, begon het bedrijf na te denken over het bouwen van gecentraliseerde ML-platforms (machine learning) met door GPU aangedreven servers in haar datacenters of in de public cloud toen de nieuwste NVIDIA® A100 GPU's beschikbaar kwamen.

Daisuke Takahashi, Solution Architect, CIU, Group IT Department bij CyberAgent, Inc. zegt: "Wij hadden voor een public cloud kunnen kiezen als we alleen GPU's wilden gebruiken, maar met een public cloud weet je nooit wanneer de nieuwste GPU's beschikbaar komen. Er kon ook niet worden gegarandeerd dat de GPU's beschikbaar zouden zijn wanneer wij ze nodig hadden, dus besloten we om gebruiksvriendelijke GPU-resources on premise in te zetten. Om de flexibiliteit van de infrastructuur heen en weer te laten gaan tussen de public cloud en de private cloud, hebben we een gebruikersinterface gemaakt die zo dicht mogelijk bij de specificaties van de public cloud ligt." CyberAgent bouwde haar eerste on-premise ML-platform met Dell PowerEdge XE8545 servers met vier NVIDIA A100 GPU's.

Waarom CyberAgent heeft gekozen voor de PowerEdge XE9680 servers met NVIDIA H100 GPU's

CyberAgent bleef de GPU-innovatie volgen, met name de nieuwste NVIDIA H100 GPU. "Wij vonden dit niet alleen aantrekkelijk vanwege de verbeterde prestaties, maar ook vanwege mechanismen zoals de Transformer Engine die specifieke rekenalgoritmen versnellen", legt dhr. Takahashi uit. "Volgens NVIDIA kan de Transformer Engine de AI-training van LLM's tot negen keer en de AI-inferentie tot 30 keer versnellen in vergelijking met de vorige generatie NVIDIA A100 GPU's."

CyberAgent koos voor het PowerEdge XE9680 servermodel met acht NVIDIA H100 GPU's. Takahashi legt verder uit: "Toen we hoorden dat de Dell PowerEdge XE9680 servers met NVIDIA H100 GPU's zouden worden uitgebracht, besloten we deze zo snel mogelijk te implementeren. Wij hadden nauw contact met Dell Technologies over de configuraties die mogelijk zouden zijn met de nieuwe PowerEdge XE9680 servers en de GPU's. Wij wilden de uptime verhogen met zo weinig mogelijk units, dus het deed ons deugd dat Dell Technologies ons tegen een redelijke prijs een hoog onderhoudsniveau inclusief vier uur on-site service kon aanbieden."

DELL Technologies

Versnelt een LLM met 13 miljard parameters momenteel met een factor 5,14 en meer dan 10 keer in de toekomst.

Takahashi vervolgt: “Wij kozen ook voor de PowerEdge XE9680 servers omdat eerdere installaties van PowerEdge XE8545 servers voor stabiele prestaties en onderhoudsgemak zorgden. Daarnaast waarderen we het gebruiksgemak van de Dell iDRAC beheertool voor veilig lokaal en extern serverbeheer."

Takahashi apprecieerde het dat toen de bestelling in maart 2023 werd geplaatst, de levering iets meer dan een maand later, medio mei, was afgerond. “Met toeleveringsketens die niet soepel verliepen door de pandemie, was het ook een geruststelling dat Dell Technologies een relatief stabiele toeleveringsketen heeft en was het prettig om te weten dat zij in zo'n korte tijd konden leveren."

Er werden na oplevering verschillende innovaties doorgevoerd in het bouwproces. Takahashi herinnert zich: “Voor een LLM met een groot aantal parameters hadden we meerdere GPU's nodig, dus installeerden we acht 400 Gbps netwerkinterfacekaarten (NIC's) in elke server en gebruikten we de RDMA-technologie (Remote Direct Memory Access) voor een zeer snelle interconnectie tussen de servers. GPU-servers genereren veel warmte, dus is het belangrijk dat ze efficiënt kunnen worden gekoeld. De 6U-vormfactor van de PowerEdge XE9680 servers voor krachtige koeling is zeker aan te bevelen. Daarnaast werd het datacenter verplaatst naar een nieuwe locatie waar warmtewisselaars aan de achterkant beschikbaar zijn, zodat effectief kon worden gekoeld door watergekoelde warmtewisselaars aan de achterkant van de racks te installeren in plaats van de hele ruimte waarin ons datacenter staat, te koelen."

De nauwkeurigheid van slogans verbeteren met Transformer Engine-optimalisaties

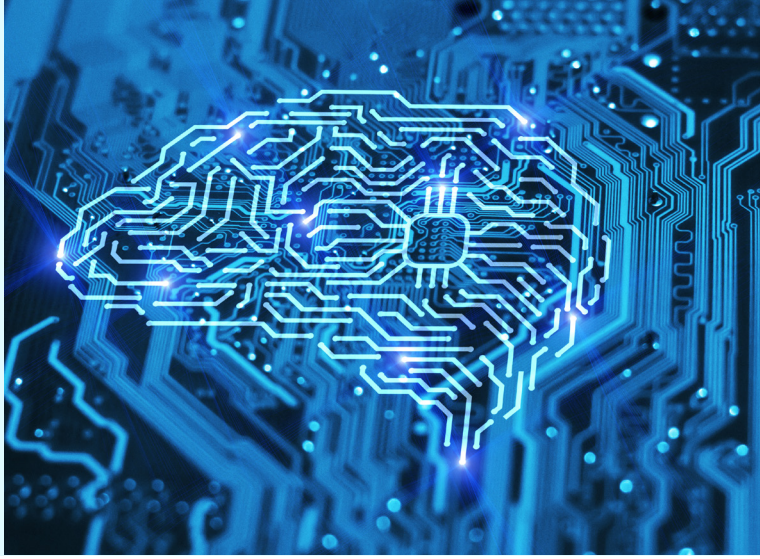
Door PowerEdge XE9680 servers te installeren, realiseert CyberAgent verschillende voordelen. “Wij verwachten dat we onze Japanse LLM's sneller en vaker kunnen bijwerken dankzij de aanzienlijke prestatieverbetering", aldus Takahashi. 'De snelheid waarmee de Japanse LLM's worden ontwikkeld, zal ook verbeteren. In vergelijking met de PowerEdge XE8545 servers die zijn uitgerust met vier NVIDIA A100 GPU's, behaalden de PowerEdge XE9680 servers met acht NVIDIA H100 GPU's bovendien een prestatieverbetering van ongeveer 5,14 keer. Wij verwachten in de toekomst ook een prestatieverhoging van meer dan 10 keer door optimalisatie voor de NVIDIA Transformer Engine. Wij kunnen voorts zeer snel ML-modellen op basis van de nieuwste datasets finetunen, waardoor we makkelijker kunnen reageren op verzoeken om onze services te ontwikkelen, de nauwkeurigheid van slogans te verbeteren en effectievere content te leveren."

De ML-infrastructuur die wordt aangedreven door PowerEdge XE9680-servers, heeft lovende kritieken van gebruikers ontvangen. “Wij hebben van onze interne onderzoekers gehoord dat ze een grotere hoeveelheid bronnen kunnen veiligstellen en gebruiken zonder zich zorgen te hoeven maken over de kosten, terwijl ze voorheen geen GPU's konden veiligstellen in de public cloud of meer moesten betalen voor langdurig gebruik”, aldus Takahashi. “Een ander voordeel is dat we een hoogwaardige infrastructuur konden leveren, inclusief interconnectie, zodat gebruikers een zakelijke impact kunnen hebben."

Takahashi is ook zeer te spreken over de beheertool iDRAC van Dell Technologies die al enige tijd in het bedrijf wordt gebruikt, omdat die tool de beheerlast vermindert. “Wij zijn niet altijd in het datacenter, dus iDRAC is handig om dingen op afstand te kunnen doen, zoals het controleren van de temperatuur en de status van de GPU's en het bijwerken van firmware zonder toegang te hoeven hebben tot het OS."

> **“De 6U-vormfactor van de PowerEdge XE9680 servers voor krachtige koeling is zeker aan te bevelen."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

> **Wij verwachten onze Japanse LLM's sneller te kunnen bijwerken. De PowerEdge XE9680 servers met acht NVIDIA H100 GPU's behaalden een prestatieverbetering van ongeveer 5,14 keer."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.



Focussen op LLM's, GPU's en infrastructuur

Met het oog op de toekomst is CyberAgent van plan om de feedback en het leerproces van OpenCALM te gebruiken om de LLM die door zijn werknemers wordt gebruikt, te verbeteren. Via OpenCALM verkent CyberAgent ook samenwerkingsverbanden met bedrijven en organisaties in andere sectoren dan de reclamewereld. CyberAgent is bijvoorbeeld gesprekken aangegaan met spelers in de detailhandel en de financiële sector om branchespecifieke LLM's te bouwen die leren van hun branchespecifieke data.

Ondertussen verklaart Takahashi dat hij op de hoogte zal blijven van de nieuwste GPU's en aanverwante nieuwe technologieën om te zien hoe die worden gecommercialiseerd. “Wij kijken er ook naar uit om te zien hoe andere leveranciers een software-ecosysteem kunnen creëren dat vergelijkbaar is met wat NVIDIA heeft bewerkstelligd. Ik ben eveneens geïnteresseerd in de implementatie van NVIDIA NVLink-C2C en nieuwe standaarden zoals CXL (Compute eXpress Link) die de CPU en GPU verbinden, omdat de PCIe-bus een bottleneck kan zijn voor de prestaties van GPU's. Ik verwacht dat Dell Technologies in hoog tempo nieuwe technologieën zal blijven omarmen en producten zal blijven ontwerpen die topprestaties leveren."

Door gebruik te maken van de nieuwste en meest kosteneffectieve GPU's zal het team van CyberAgent voor onderzoek naar AI en het ontwikkelen van AI zich blijven ontwikkelen door de ML-infrastructuur te bieden waar gebruikers om vragen. Bovendien zal CyberAgent met de verdere ontwikkeling van de Japanse LLM aanzienlijk de aandacht blijven trekken, niet alleen in de eigen reclamebusiness maar ook op de Japanse AI-markt.

> **Wij wilden de uptime verhogen met zo weinig mogelijk units, dus het deed ons deugd dat Dell Technologies ons tegen een redelijke prijs een hoog onderhoudsniveau inclusief vier uur on-site service kon aanbieden."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department,
> CyberAgent, Inc.

Deze inhoud is door Dell Technologies vanuit de Japanse versie vertaald.

Meer informatie over Generative AI-oplossingen van Dell Technologies.

Maak verbinding op social media.

DELL Technologies

Copyright © 2023 Dell Inc. of zijn dochterondernemingen. Alle rechten voorbehouden. Dell Technologies, Dell en andere handelsmerken zijn handelsmerken van Dell Inc. of haar dochterondernemingen. Andere handelsmerken zijn mogelijk handelsmerken van hun respectieve eigenaren. Deze casestudy is alleen voor informatiedoeleinden. Dell is van mening dat de informatie in deze casestudy correct is op de publicatiedatum, september 2023. De informatie kan zonder voorafgaande kennisgeving worden gewijzigd. Dell geeft geen garanties, expliciet noch impliciet, in deze casestudy.