DELL Technologies

MEDIA AND ENTERTAINMENT
JAPAN

CyberAgent

Developing Japanese Generative AI and transforming digital advertising services

CyberAgent, Inc. uses Dell PowerEdge XE9680 Servers with eight NVIDIA® H100 Tensor Core GPUs to accelerate Generative AI and improve advertising effectiveness.

Business needs

Since 2016, CyberAgent, Inc. has been actively researching and developing AI, and incorporating it into its advertising business. The company needed to provide its staff with fast, affordable access to highly reliable on-premises servers with the most advanced NVIDIA GPUs available for its Generative AI development efforts.

Business results

- Accelerates large language model (LLM) performance by approximately 5.14x compared to the previous generation with PowerEdge XE9680 servers.
- Expects performance improvement by more than 10x in the future with NVIDIA Transformer Engine optimizations.
- Enables high-speed fine-tuning of machine learning models according to the latest datasets.
- Saves data center space and delivers efficient cooling with a 6U form factor versus the mainstream 8U.

Solutions at a glance

- Dell PowerEdge XE9680 servers with NVIDIA® H100 GPUs
- Dell ProSupport

CyberAgent, Inc. is a company known for being the market leader in the domestic internet advertising industry and ventures including the innovative TV platform, ABEMA. In 2016, the company established an AI research organization called AI Lab and has since been actively researching and developing AI. In 2020, CyberAgent introduced a cutting-edge predictive AI that improves the production of high-impact banner ad catchphrases and image combinations, boosting advertising effectiveness.

CyberAgent continued its Generative AI development, creating a unique Japanese language–specific large language model (LLM) with 13 billion parameters. Designed as a general-purpose AI model that can be used in a variety of situations, the LLM can be fine-tuned to create catchphrase copies that resonate with the users of each advertising platform. CyberAgent is already using its Japanese LLM in AI services such as Kiwami Prediction AI, Kiwami Prediction TD and Kiwami Prediction LP to support creative-advertising production and predict advertising effectiveness. In the future, CyberAgent aims to develop a multimodal AI that can handle not only Japanese LLMs but also images.

> **"Our in-house researchers can secure a larger amount of resources and use them without worrying about the cost, whereas previously they could not secure GPUs in the public cloud or were charged more for long-term use."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

In May 2023, CyberAgent released a commercially available open-source Japanese LLM called OpenCALM (Open CyberAgent Language Models), which includes up to 6.8 billion parameters.

While ChatGPT is tuned for chatting, OpenCALM is more of a general-purpose Japanese language model that can be fine-tuned to suit users' needs. CyberAgent released OpenCALM as an open-source project because it is more beneficial for the company to receive feedback from other sources and collaborate with other companies to contribute to the development of AI technology in Japan, rather than developing a Japanese LLM in a closed environment.

The infrastructure powering CyberAgent's AI innovation

When CyberAgent established its AI Lab in 2016, each researcher had a GPU-powered workstation for research. However, the necessity of remote work during the 2020 pandemic made it difficult for each researcher to utilize GPU-powered workstations. To ensure researchers had the computing resources they needed, the company began thinking about building centralized machine learning (ML) platforms with GPU-powered servers either in its data centers or in the public cloud when the latest NVIDIA® A100 GPUs were released.

Daisuke Takahashi, Solution Architect, CIU, Group IT Department at CyberAgent, Inc., says, "We could have chosen a public cloud if we just wanted to use GPUs, but with a public cloud, you never know when the latest GPUs will become available. Also, there's no guarantee that the GPUs would be available when we wanted them, so we decided to deploy GPU resources on premises with ease of usage. To realize the flexibility of the infrastructure to move back and forth between the public cloud and private cloud, we devised a user interface that is as close to the public cloud specifications as possible." CyberAgent built its initial on-premises ML platform using Dell PowerEdge XE8545 servers with four NVIDIA A100 GPUs.

Why CyberAgent selected the PowerEdge XE9680 Servers with NVIDIA H100 GPUs

CyberAgent continued to follow GPU innovation, especially the latest NVIDIA H100 GPU. "We thought it was attractive not simply for its improved performance, but also mechanisms such as its Transformer Engine that accelerate specific computational algorithms," Mr. Takahashi explains. "According to NVIDIA, the Transformer Engine can accelerate AI training of LLMs by up to nine times and AI inference by up to 30 times compared to the previous-generation NVIDIA A100 GPUs."

CyberAgent chose the PowerEdge XE9680 server model with eight NVIDIA H100 GPUs. Mr. Takahashi explains, "When we learned that the Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs would be released, we decided to adopt it as soon as possible. We were able to communicate closely with Dell Technologies about the configurations that would be possible with the upcoming PowerEdge XE9680 servers and the GPUs. We wanted to increase uptime with as few units as possible, so we were pleased that Dell Technologies was able to provide us with a high level of maintenance, including four-hour on-site service, at a reasonable price."

DELL Technologies

Accelerates an LLM with 13 billion parameters by 5.14x today and more than 10x in the future.

Mr. Takahashi continues, “We also chose the PowerEdge XE9680 servers because previous installations of PowerEdge XE8545 servers have provided stable performance and ease of maintenance. In addition, we value the ease of use of the Dell iDRAC management tool for secure local and remote server management."

Mr. Takahashi appreciates the fact that when the order was placed in March 2023, delivery was completed a little over a month thereafter, in mid-May. “With supply chains in disarray due to the pandemic, I was also reassured that Dell Technologies has a relatively stable supply chain, and it was nice to know that they could deliver in such a short period of time."

A variety of innovations were made in the post-delivery build process. Mr. Takahashi recalls, “For a LLM with a large number of parameters, we needed to use multiple GPUs, so we installed eight 400Gbps network interface cards (NICs) in each server and used RDMA (Remote Direct Memory Access) technology to create a high-speed interconnect between the servers. GPU servers generate a lot of heat, so it is important that they are designed to be cooled efficiently. The PowerEdge XE9680 servers 6U form factor for solid cooling is also commendable. In addition to that, the data center was also relocated to a new location where rear door heat exchangers are available so that effective cooling can be achieved by installing water-cooled, rear door heat exchangers at the rear of the racks, rather than cooling the entire room that houses our data center."

Improving catchphrases accuracy with Transformer Engine optimizations

By installing PowerEdge XE9680 servers, CyberAgent is realizing a variety of benefits. “We expect to be able to update our Japanese LLMs faster and more frequently due to the significant improvement in performance," says Mr. Takahashi. “The speed of evolution of Japanese LLMs will also improve. Moreover, compared to the PowerEdge XE8545 servers equipped with four NVIDIA A100 GPUs, the PowerEdge XE9680 servers with eight NVIDIA H100 GPUs achieved a performance improvement of approximately 5.14 times. We also anticipate a performance increase of more than 10 times by optimizing for the NVIDIA Transformer Engine in the future. We're also able to perform high-speed fine-tuning of ML models according to the latest datasets, which will make it easier to respond to requests to evolve our services, improve the accuracy of catchphrases and deliver more effective content."

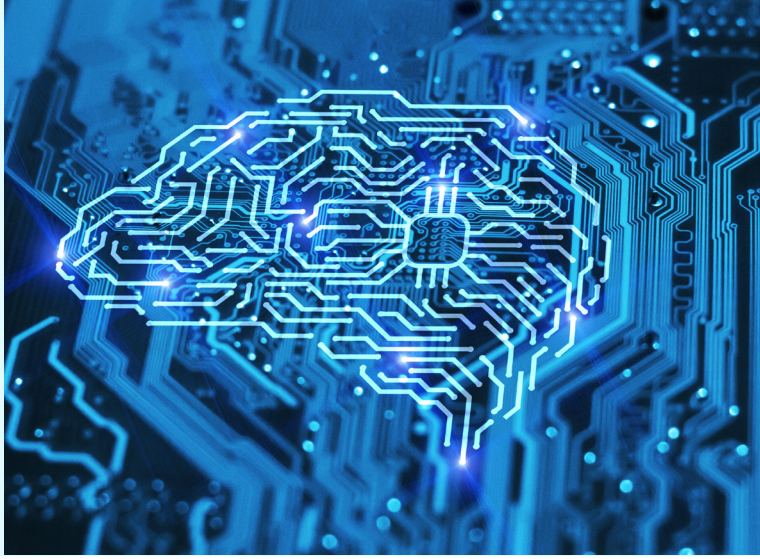
The ML infrastructure powered by PowerEdge XE9680 servers has received high praise from users. “We have heard from our in-house researchers that they can secure a larger amount of resources and use them without worrying about the cost, whereas previously they could not secure GPUs in the public cloud or were charged more for long-term use," says Mr. Takahashi. “Another benefit is that we were able to provide a high-spec infrastructure, including interconnect, so that users can make a business impact."

Mr. Takahashi also appreciates the Dell Technologies iDRAC management tool, which the company has been using for some time, because it reduces management burden. “We are not always stationed in the data center, so iDRAC is useful for doing things remotely, such as checking the temperature and status of the GPUs and updating firmware without having to access the OS."

> **“The PowerEdge XE9680 servers 6U form factor for solid cooling is also commendable."**
>
> **Daisuke Takahashi**
> Solution Architect, CIU, Group IT Department, CyberAgent, Inc.

> **We expect to be able to update our Japanese LLMs faster. The PowerEdge XE9680 servers with eight NVIDIA H100 GPUs achieved a performance improvement of approximately 5.14 times."**

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



Focusing on LLMs, GPUs and infrastructure

Looking ahead, CyberAgent plans to utilize the feedback and learning gathered from OpenCALM to improve the LLM that its employees are using. Through OpenCALM, CyberAgent is also exploring collaborations with companies and organizations in industries other than advertising. For example, CyberAgent has begun discussions with players in retail and finance to build industry-specific LLMs that learn from their industry-specific data.

Meanwhile, Mr. Takahashi explains that he will continue to be updated with the latest GPUs and related new technologies to see how they are commercialized. "We are also looking forward to seeing how other vendors can create a software ecosystem similar to the one NVIDIA has achieved. I am also interested in the implementation of NVIDIA NVLink-C2C and new standards such as CXL (Compute eXpress Link) that connect the CPU and GPU, as the PCIe bus can be a bottleneck to GPU performance. I expect Dell Technologies to continue to adopt new technologies at a rapid pace and design products that deliver on performance."

By using the latest and cost-effective GPUs, CyberAgent's AI research and development team will continue to evolve by providing the ML infrastructure that users demand. In addition, with the further development of Japanese LLM, CyberAgent will continue to attract significant attention, not only in its own advertising business but also in the Japanese AI market.

> **We wanted to increase uptime with as few units as possible, so we were pleased that Dell Technologies was able to provide us with a high level of maintenance, including four-hour on-site service, at a reasonable price."**

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

This content has been translated from the Japanese version by Dell Technologies.