CUSTOMER SUCCESS STORY

CyberAgent accelerates AI with a hybrid cloud

NetApp

The choice of NetApp AFF A800 and Trident for increases cost performance by a factor of 4.

CyberAgent has promoted a growth strategy based around online advertising, gaming, and media. The company is known for its innovative AI-powered services for retail and healthcare, particularly digital advertising. In May 2020, CyberAgent shocked the industry by launching a predictive AI system that applies what it calls “Kiwami Prediction AI” to the production of advertising creatives.

To further accelerate AI development at a company-wide level, CyberAgent’s AI Business Unit has adopted the latest GPU server NVIDIA DGX A100 and the high-end NetApp AFF A800 All-Flash Array to construct a new AI learning platform using containers and Kubernetes. By linking cloud and on-premises environments, this platform provides twice the performance of a conventional cloud service at half the cost.

CyberAgent

4x

increase in cost performance

"Using Trident enables dynamic storage provisioning from Kubernetes, so users don't need to worry about storage, greatly improving the usability of our AI learning platform. When we found out how effective Trident could be, we knew it just had to be NetApp."

Masaya Aoyama
Infrastructure Engineer/Developer Expert (Kubernetes/Cloud Native), AI Business Unit
CyberAgent, Inc.

AI-powered digital advertising = higher engagement

CyberAgent AI strategy is demonstrated by the development and commercialization of predictive AI, applying AI to produce advertising creatives. Making this a performance-based service, where production costs are only charged when the advertising proves to be effective, has demonstrated the power of AI technology to advertisers and the rest of the industry.

"In the rapidly evolving world of digital advertising, it's no exaggeration to say that the competitiveness of your business depends on how well you leverage AI," said Lee Yeongjae, AI Platform Project Manager.

"We formed the AI Business Unit in September 2019 to accelerate the development of digital marketing services leveraging AI. We belong to the AI Tech Studio, a development group of about 250 engineers, working on research and development of AI technologies and their application and implementation."

CyberAgent built an AI learning platform on Google Cloud Platform (GCP), and has been using it to collect learning data and develop AI models. But increasing development demand together with rising cloud costs posed a serious problem.

"It was spring 2020 when we started considering a new AI learning platform to respond to rapidly increasing AI development needs while keeping costs down. Our basic policy was to build an on-premises AI learning platform that could fully leverage the latest GPUs, to significantly improve cost performance," said Lee Yeongjae.

There was another requirement for an AI learning platform allowing multiple engineers to run learning tasks at the same time: it must support containers and Kubernetes, said Masaya Aoyama, AI Business Unit Infrastructure Engineer/Developer Expert.

"Our aim was to create a multi-tenancy AI learning platform, leveraging the portability and reproducibility of containers," Masaya Aoyama said. "We selected NetApp's AFF A800 All-Flash Array as the storage product offering the best compatibility with Kubernetes and the Container Storage Interface (CSI)."

Storage Orchestrator Trident supports CSI

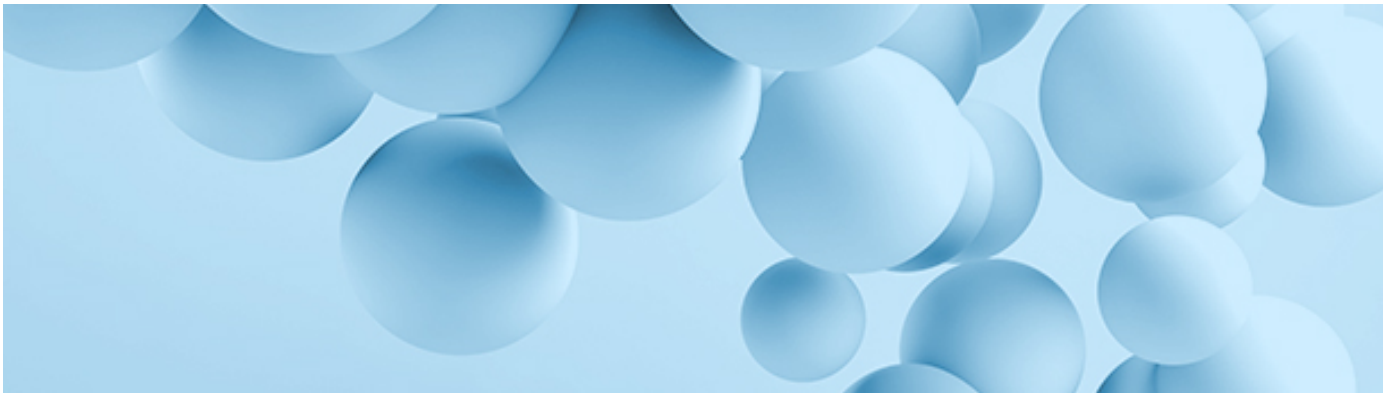
The NetApp AFF A Series All-Flash Array boasts industry-leading performance and the outstanding features of NetApp ONTAP data management software. For their AI learning platform, CyberAgent adopted the NetApp AFF A800 with NVMe SSD, which uses 4U active-active dual controllers for exceptional IOPS performance and multiple processing capability. Aoyama, Lee, and Takahashi set out three requirements for the new AI learning platform.

(1) A high-performance on-premises AI learning platform capable of fully leveraging the latest GPU
(2) A superior multi-tenancy AI learning platform utilizing containers
(3) An environment that can easily handle persistent storage from Kubernetes via CSI

"First, we decided on the NVIDIA DGX A100 as the GPU server. We wanted to be the first to introduce the highest performance GPU to enhance our AI development capacity. To get the most from the GPU, high-performance storage is essential to provide the learning data. We shortlisted three possible storage products and verified them with the NVIDIA DGX reference architecture," said Daisuke Takahashi of the AI Business Unit.

The team were impressed by the stability of NetApp's AFF A800 All-Flash Array when transferring data at full bandwidth.

"The deciding factor was the compatibility with Kubernetes and CSI," Masaya Aoyama said "These software solutions are open source community-developed and are constantly evolving. It must be



difficult for storage vendors to keep up with this speed and continually update CSI drivers, but NetApp's Trident had far and away the best compatibility."

Available since 2016, Trident is a CSI driver that integrates with Kubernetes to provision persistent volume requests from NetApp storage platforms, offering outstanding functionality as a storage orchestrator and provisioner. Trident is developed and updated by an open source project led by NetApp, and is available free of charge.

"Using Trident enables dynamic storage provisioning from Kubernetes, so users don't need to worry about storage, greatly improving the usability of our AI learning platform. When we found out how effective Trident could be, we knew it just had to be NetApp," said Aoyama.

A superior multi-tenancy AI learning platform

Masaya Aoyama is an expert in Kubernetes and Cloud Native development. He is well known in the OSS community, frequently giving talks and writing articles.

"Using containers frees engineers from the hassle of setting up the learning environment," he said "AI learning involves all kinds of tools and library versions, so it's practically impossible to manually ensure conformity or recreate an environment used previously. But these issues can be solved simply by packaging the implementation environment using containers."

The beauty of Kubernetes is that it automates the tasks involved in operating containers. In this environment, a scheduler automatically assigns the best Node (GPU and storage resources) to a Pod (unit of containers), helping to improve productivity when developing AI learning models in a multi-tenant environment.

"The NetApp technical team responded quickly to any questions or issues that arose during the process of verification and introduction,"said Lee. "We were able to resolve most of these at our meetings, but even for matters that required detailed confirmation, they usually got back to us the same day. For questions relating to Trident, they explained by linking us to the community, which deepened our knowledge."

"Looking in at the Trident community was sometimes just like seeing how our internal team works. The members of the AI Tech Studio are used to sharing information and managing tasks using GitHub, so we have a similar culture," said Aoyama.

Going on-premises for twice the performance at half the cost

The new AI learning platform was launched in December 2020. "It has yet to be evaluated in a real environment, but there's no doubt that it can significantly improve cost performance," said Lee.

"NVIDIA A100 boasts twice the performance of GPUs used on GCP. The cost is forecast over the next five years to be halved, so the new AI learning platform will achieve four times the cost performance," said Lee.

CyberAgent uses cloud services like GCP and AWS and on-premises environments for different purposes, but until now, AI development and implementation has mainly been done in a cloud environment.

"The new environment allows learning data to be transferred from GCP or AWS to the on-premises AI learning platform as required, facilitating AI learning with maximum performance. To streamline this procedure, we're considering using NetApp's Cloud Volumes ONTAP on GCP so that data can be moved to and from the on-premises NetApp AFF storage," Lee said.

Cloud Volumes ONTAP (CVO) is a solution allowing the outstanding features of ONTAP data management software, like SnapMirror (replication), FlexClone (cloning), data deduplication and compression, to be used on a cloud platform like GCP or AWS. "This enables data to be stored and leveraged optimally depending on the purpose, taking into account return on investment and latency. We want to promote the use of CVO and SnapMirror to get closer to an ideal hybrid environment linking cloud and on-premises environments," Lee said.

Evolving at the speed of the cloud

The on-premises AI learning platform incorporates a layer compatible with GCP's AI platform, with the aim of linking learning data and processes more effectively.

"One of our main goals was to create an environment with object storage on the cloud as a hub, allowing users to freely select the learning system and where to deploy it," Lee Yeongjae said. "But in the world of digital advertising, where responses are required in a matter of milliseconds, we need to carefully consider where data and systems are stored. There are plenty of challenges, but we are now ready to move forward towards this goal."

Aoyama, who is working on CyberAgent's platform alongside his work at the forefront of cloud native solutions, concludes:

"Going forward, AI Tech Studio will be challenged with developing AI applications with even greater business impact. We need to further evolve the AI learning platform to support this. As well as keeping up with GCP updates, we want to further enhance integration of cloud and on-premises environments using Kubernetes. I hope NetApp will continue to support our business and platform as it evolves at the speed of the cloud."

Products
AFF A800
Trident
Cloud Volumes ONTAP

Protocols
NFS
Iscsi

+1 877 263 8277

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services and applications to the right people—anytime, anywhere. To learn more, visit www.netapp.com