November 22, 2024

To All Members of the Press

GMO Internet Group, Inc.

GMO Internet Group Launches “GMO GPU Cloud,” Ranked in the TOP500 Supercomputer List ～ Contributing to the Rapid Advancement of Japan’s AI Industry with One of the Fastest AI Development Environments in the Country ～

GMO Internet Group, Inc. (Group CEO: Masatoshi Kumagai, hereinafter "GMO Internet Group") has launched its high-performance GPU cloud service, “GMO GPU Cloud,” on Friday, November 22, 2024. The service has earned a spot for the first time in the global supercomputer ranking “TOP500.”

This service utilizes the high-performance “NVIDIA H200 Tensor Core GPU” (hereinafter “H200 GPU”) and the world’s first Ethernet networking platform dedicated to AI workloads, “NVIDIA Spectrum-X,” to dramatically improve the efficiency of AI development and machine learning. By doing so, it contributes to strengthening the international competitiveness of Japan’s AI industry.

【Service Overview】

・Service Name: GMO GPU Cloud
・Launch Date: Friday, November 22, 2024
・URL: https://gpucloud.gmo/

GMO GPUクラウド

【Background of the Service Launch】

With the rapid evolution of AI technologies—especially in the development of large language models (LLMs)—the demand for computing resources is skyrocketing. To accelerate AI research and development in Japan and enhance its international competitiveness, a world-class computing environment is essential. GMO Internet Group has decided to launch the “GMO GPU Cloud” in response to this national imperative.

【About “GMO GPU Cloud”】（URL : https://gpucloud.gmo/）

“GMO GPU Cloud” is one of the fastest GPU cloud services available in Japan. By utilizing the high-performance “H200 GPU,” it significantly reduces training time for large language models, dramatically boosting the efficiency of AI development.

Furthermore, GMO Internet Group is the first domestic cloud provider to adopt both the “H200 GPU” and the “Spectrum-X” Ethernet networking platform. This enables the provision of top-tier GPUs connected via Spectrum-X networking. The combination of “H200 GPU” and “NVIDIA Spectrum-X” delivers a highly optimized GPU cloud environment for generative AI development and machine learning.

Through this service, GMO Internet Group offers a finely-tuned, high-performance computing environment—requiring no infrastructure adjustments—to businesses and research institutions engaged in generative AI and high-performance computing (HPC). This contributes to shorter development timelines and reduced costs, while promoting the advancement of Japan’s AI industry.

■ Key Features of “GMO GPU Cloud”

1. Equipped with “NVIDIA H200 Tensor Core GPU”

The H200 GPU is designed for developers and researchers working with large language models, with significantly expanded GPU memory capacity and memory bus bandwidth. It offers approximately 1.7 times the memory capacity and about 1.4 times the bandwidth of the NVIDIA H100 Tensor Core.

▲NVIDIA H200 Tensor Core GPU

2. First Domestic Cloud Provider to Adopt “NVIDIA Spectrum-X”

GMO GPU Cloud is the first in Japan to adopt Spectrum-X, a networking platform that dramatically enhances Ethernet performance for AI workloads.

▲NVIDIA Spectrum-X

3. Cloud Network Acceleration with NVIDIA BlueField-3 DPU

The NVIDIA BlueField-3 data processing unit accelerates GPU access to data, streamlines the delivery of AI applications, and enhances cloud infrastructure security.

4. Ultra-High-Speed Storage by DDN

▲DDN High-Speed Storage

Utilizing DDN’s high-speed storage optimized for integration with the NVIDIA platform, the service offers a powerful, all-in-one AI development platform.

5. Fast Environment Setup and Management with NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native software platform that accelerates data science pipelines and streamlines the development and deployment of production-grade copilots and other generative AI applications.

▲NVIDIA AI Enterprise

6. Industry-Standard Slurm Job Scheduler

Slurm, the industry-standard job scheduler for cluster systems, is used to manage resource allocation, job control, and monitoring.

▲ Slurm worked manager

■ Pricing (excluding tax):

	Dedicated Plan	Shared Plan
Usage Fees	¥3.8 million/month per server	GPU side: ¥100/min CPU side: ¥20/min
Base Fee	-	50% of monthly contract (applied to usage fees)
Maximum Number of Users	50users	10users
Local Storage	15TB per server free (as temporary storage during job submission)	
Home Storage	100GB per user free	
High-Speed Shared Storage	¥30,000/TB per month (for 1TB to 100TB)	

■Use Cases

- High-speed training and fine-tuning of large language models
- Training computer vision models using large-scale datasets
- Scientific computing for applications such as drug discovery and weather forecasting
- Research and development requiring high-performance computing (HPC)

<Comment from Masatoshi Kumagai, Group CEO and President>

At GMO Internet Group, we have viewed AI as a transformative force on par with the Internet revolution, and have been actively focusing on its potential and investing in research and development for over a decade, starting around ten years ago.

Today, we are proud to officially launch “GMO GPU Cloud,” a GPU cloud service that boasts one of the fastest performance levels in Japan and has secured a spot in the TOP500 global supercomputer rankings.

This service dramatically enhances the speed and efficiency of research and development across a wide range of fields that require advanced computing resources—from high-speed training of large language models, to the development of computer vision models, and even scientific computing for applications such as drug discovery and weather forecasting.

Leveraging our 30 years of experience and proven track record in supporting Japan’s Internet infrastructure, we are committed to delivering advanced computing environments—the foundation for AI development—to as many researchers and developers as possible.

【Ranked in the “TOP500”】

On November 18, 2024, “GMO GPU Cloud” was ranked 37th in the world and 6th in Japan in the TOP500 list of supercomputer performance.Among commercial cloud services in Japan, it achieved the No.1 ranking (※1).

Built with 96 nodes (768 GPUs) using the latest “H200 GPU” (※2), it recorded a LINPACK performance (※3) of 38.06 PFLOPS (petaflops) with an execution efficiency of 73.0%.

(Reference) GMO Internet Group’s “GMO GPU Cloud” makes its first appearance in the global TOP500 supercomputer ranking: https://www.gmo.jp/news/article/9266/

(※1) According to our research, as of November 22, 2024, among commercial cloud services available in Japan

(※2) The 96-node configuration refers to a system structure in which 96 servers (768 GPUs) are interconnected to perform parallel processing.

(※3) LINPACK performance is one of the standard benchmarks used to measure the computing power of supercomputers and high-performance computing systems.

【Future Outlook】

GMO Internet Group will continue to contribute to the advancement of this new era by providing cutting-edge infrastructure, centered on “GMO GPU Cloud,” in line with the progress of the AI and robotics revolution.

Our goal is to become widely recognized as “the go-to for GPU cloud services,” delivering new value to the market as an essential cloud solution for the AI industry.

■Past Reference Releases

• April 19, 2024: GMO Internet Group to become the first in Japan to offer the fastest GPU cloud service for generative AI using NVIDIA H200 Tensor Core GPUs
https://www.gmo.jp/news/article/8933/

• June 11, 2024: GMO Internet Group becomes the first Japanese cloud provider to adopt NVIDIA Spectrum-X for its generative AI GPU cloud service
https://www.gmo.jp/news/article/9005/

• August 29, 2024: GMO Internet Group releases “GPU Cloud Usage Survey” – Domestic usage rate at only 5.4%, with nearly 90% using overseas services
https://www.gmo.jp/news/article/9133/

• September 26, 2024: GMO Internet Group demonstrates the performance of its environment equipped with “NVIDIA H200 GPU”
https://www.gmo.jp/news/article/9164/

• November 13, 2024: GMO Internet Group discusses infrastructure and security for the AI and robotics era at “NVIDIA AI Summit”
https://www.gmo.jp/news/article/9233/

• November 19, 2024: GMO Internet Group’s “GMO GPU Cloud” makes its debut on the global TOP500 supercomputer ranking
https://www.gmo.jp/news/article/9266/

【About GMO Internet Group, Inc.】

Since launching its Internet business in December 1995, GMO Internet Group, Inc. has operated under the corporate slogan “Internet for Everyone,” focusing its management resources on creating and expanding the Internet space to make the Internet more accessible and enriching. Today, the company operates in four main segments: Internet Infrastructure, Online Advertising and Media, Internet Finance, and Crypto assets. As of the end of September 2024, the Group serves 15.18 million customers and consists of 111 companies, including 10 publicly listed firms, with approximately 7,500 partners—growing into a comprehensive Internet group.
Under the vision “To become the No.1 corporate group that creates the future with AI,” we are fully embracing generative AI across the Group to:
① Save time and reduce costs
② Enhance the quality of existing services
③ Develop new services for the AI industry.

Press Inquiry
GMO Internet Group, Inc.
Kawaberi,Public Relations
Corporate Business Management Division,
TEL : +81-3-5456-5555　Contact:pr@gmo.jp

GMO Internet Group, Inc.
Yamazaki, PR Team,
Group Corporate Communications Department
TEL: +81-3-5456-2695
Contact:https://www.gmo.jp/contact/press-inquiries/

Service Inquiry
GMO Internet Group, Inc.
Domain & Hosting Business Division
Contact:aicloud@gmo.jp

[GMO Internet Group, Inc.] (URL: https://www.gmo.jp/)

Company Name	GMO Internet Group, Inc. (TSE Prime Market Securities Code: 9449)
Location	Cerulean Tower 26-1 Sakuragaoka-cho, Shibuya-ku, Tokyo
Representative	Masatoshi Kumagai, Founder, Chairman and Group CEO
Business	Internet Infrastructure Online Advertising and Media Internet Finance Crypto assets
Capital stock	5 billion yen