PRESS RELEASE

December 16, 2025

FOR IMMEDIATE RELEASEG

GMO Internet, Inc.

GMO GPU Cloud Launches one of Japan's First NVIDIA HGX B300 Cloud Services

—Powered by NVIDIA Blackwell Ultra GPU, Optimized for AI Inference—

GMO Internet, Inc. (Representative Director, President and CEO: Tadashi Ito), part of GMO Internet Group, today launched a cloud service featuring NVIDIA HGX B300 AI infrastructure (hereinafter "HGX B300") powered by NVIDIA Blackwell Ultra GPUs on its high-performance GPU cloud service "GMO GPU Cloud," built on NVIDIA technology. This marks one of the fastest launches in Japan.

Through a bare metal configuration with dedicated server access, the service maximizes the performance of HGX B300 AI infrastructure, which is optimized for AI inference models. It provides a high-performance environment ideal for cutting-edge AI workloads including high-speed inference of large language models and agentic AI development, all delivered from within Japan. GMO Internet will deploy this as a new computing infrastructure supporting not only R&D for AI and robotics,but also practical AI implementation.

(*1) According to company research as of December 22, 2025

Background

In recent years, demand for large-scale computing resources capable of advanced computational processing has been growing rapidly, driven by corporate development of proprietary LLMs (large language models) and the rapid evolution of AI and robotics in industrial sectors. To meet requirements such as the expansion of AI models, real-time inference, and support for diverse generation formats, cloud infrastructure must continuously evolve as a "dynamic technology foundation," with regular updates being key to technological innovation.

GMO Internet received certification from the Ministry of Economy, Trade and Industry on April 15, 2024, for its supply security plan for "cloud programs," a specified critical material under the Economic Security Promotion Act (*1). Since November 22 of the same year, the company has been providing NVIDIA Hopper GPUs at industry-leading speeds in Japan (*2). In response to market conditions and technical requirements, we reviewed our additional investment plan announced on May 14, 2025, and decided to deploy the latest generation of the Blackwell architecture, "HGX B300."

This enables "GMO GPU Cloud" to support the high computational performance, scalability, and low-latency communication required for future AI development and operations, evolving as an even more advanced infrastructure platform.

(*1) Ministry of Economy, Trade and Industry: Ensuring Stable Supply of Specified Critical Materials under the Economic Security Promotion Act https://www.meti.go.jp/policy/economy/economic_security/cloud/index.html

(*2) Press Release: "GMO Internet Group Launches 'GMO GPU Cloud' Ranked in TOP500 Supercomputer List" https://internet.gmo/news/article/27/

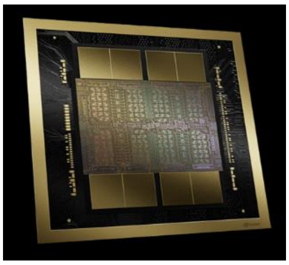
GMO GPUクラウド

About GMO GPU Cloud (NVIDIA B300)

1. GMO GPU Cloud Overview

Service Name	GMO GPU Cloud Bare Metal Plan
Launch Date	December 16, 2025
GPU Configuration	NVIDIA Blackwell Ultra GPU
GPU Memory	2.3 TB
CPU Configuration	Intel Xeon 6767P x2
Main Memory	3 TB
Local Storage	3.84TB NVMe Gen5 SSD x8
Network Storage	NFS File Storage and Object Storage (Coming Soon)
Global/Local Storage Network (Shared)	NVIDIA BlueField-3 DPU 200Gbps x2
Interconnect	NVIDIA Connect-X 8*8 total 6,400Gbps
Use Cases	- High-speed training and fine-tuning of large language models - Accelerated AI reasoning and inference processing - Training computer vision models with large-scale datasets-
Pricing	Custom quotation per company

2. NVIDIA Blackwell Ultra GPU (GPU Memory) Features



Specifications (vs. NVIDIA H200)

Memory capacity: 288GB (204% of H200)

Newly supports FP4 computation precision

FP8 Tensor Core performance improved up to 2.25x

3. NVIDIA HGX B300 Features

Enhanced Efficiency for Large-Scale Training

・Training performance improved up to 4x (vs. previous generation H200)
・Key performance metrics including GPU memory capacity and bandwidth significantly enhanced
・Improved performance per watt

Strengthened Inference Capabilities

・Inference performance improved 11x (vs. previous generation H200)
・FP4 support reduces VRAM usage while improving throughput
・Largest GPU memory capacity to date

4. Network Configuration

NVIDIA HGX B300 AI infrastructure is provided with ultra-high-speed communication with up to 6 400Gbps NVIDIA ConnectX-8 SuperNICs and NVIDIA Spectrum-X Ethernet switches that connect and scale multiple GPUs. Its low-latency design accelerates data transfer during large-scale AI training and delivers stable high performance even in distributed learning environments combining multiple servers.

NVIDIA ConnectX-8

NVIDIA BlueField3-DPU

NVIDIA Spectrum-4

5. NVIDIA B300 と NVIDIA H200 の比較

	NVIDIA HGX B300	NVIDIA HGX H200
GPU Memory	計 2.3 TB（270GB HBMe3 x8）	計 1.1 TB（141GB HBMe3 x8）
FP4 Tensor Core	144PFLOPS/108PFLOPS	-

FP8 Tensor Core	72PFLOPS	32PFLOPS
FP16 Tensor Core	36PFLOPS	16PFLOPS
FP32	600TFLOPS	540TFLOPS
Max Interconnect Bandwidth	6,400Gbps	3,200Gbps
NVLINK (Inter-GPU Bandwidth/Per GPU)	5th Gen (1,800GB/s)	4th Gen (900GB/s)

About the Bare Metal Plan

The "GMO GPU Cloud Bare Metal Plan," launching as Japan's first offering today, provides a fully dedicated environment that maximizes computational performance by accessing GPUs directly without a virtualization layer. It leverages computational capabilities specialized for AI inference processing and is suitable for diverse workloads including high-speed inference of large language models and agent AI development.

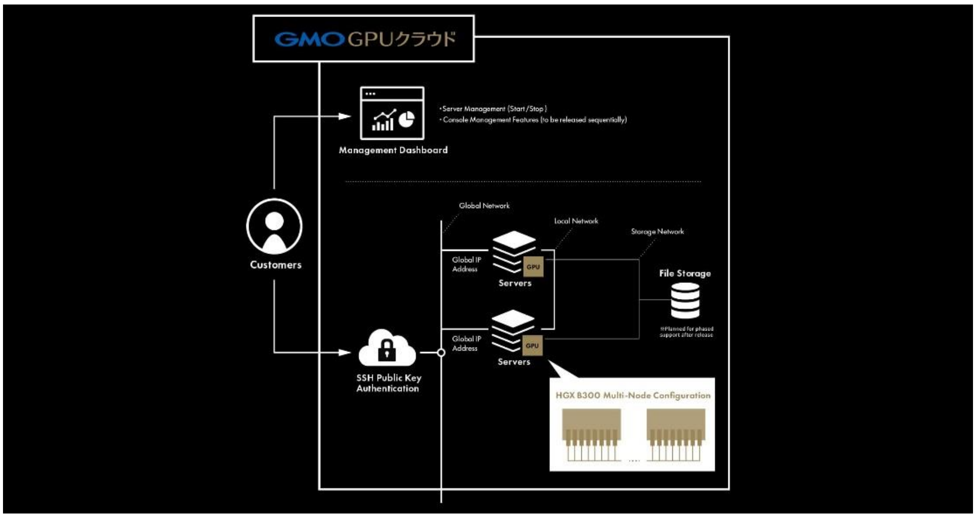
The service also maintains the flexibility inherent to bare metal while incorporating mechanisms to reduce the effort required for initial setup and network construction.

■**Key Features:**

- Provided with Ubuntu OS pre-configured
- Pre-configured interconnect network (local network also available)
- Server management via management dashboard
- Additional storage available

This enables users to utilize a GPU environment that balances high performance with flexible tuning capabilities while minimizing deployment effort.

Server Configuration Image for NVIDIA B300


GMO GPUクラウド
・Server Management (Start/Stop)
・Console Management Features (to be released sequentially)
Management Dashboard
Customers
Global Network
Local Network
Storage Network
Global IP Address
GPU
Servers
File Storage
※Planned for phased support after release
Global IP Address
GPU
Servers
SSH Public Key Authentication
HGX B300 Multi-Node Configuration

Comment from Tadashi Ito, Representative Director, President and CEO

Since the launch of GMO GPU Cloud in November 2024, we have been supporting infrastructure for advanced AI development. We have consistently maintained our commitment to cutting-edge, high-performance technology, stable operational environments, and comprehensive technical support, continuously working to support our customers' AI development initiatives. Being the first in Japan to provide NVIDIA B300 represents a significant step in accelerating Japan-originated AI innovation to the next level. Going forward, we will strengthen our technical collaboration with NVIDIA, enhance our AI development platform to support engineers' endeavors, and robustly promote the creation of globally competitive next-generation AI.

Comment from Masataka Osaki, Japan Country Manager, VP Worldwide Field Operations, NVIDIA

"For developers and researchers tackling AI development daily, access to vast computational resources through high-performance and scalable computing environments is essential. The launch of 'GMO GPU Cloud' equipped with NVIDIA HGX B300, which enables highly efficient training and high-speed inference, represents a significant step in leading Japan's AI infrastructure..."

Future Development

GMO Internet will contribute to technological innovation in the rapidly evolving AI and robotics fields through its AI infrastructure strategy centered on "GMO GPU Cloud." By continuing to provide the latest AI computing infrastructure and constructing flexible cloud environments tailored to customer needs, we will serve as an essential domestic AI infrastructure for Japan's AI industry, contributing to AI innovation creation for society and industry.

Background on GMO GPU Cloud (URL: https://gpucloud.gmo/)

"GMO GPU Cloud," Japan's fastest-class GPU cloud service built on NVIDIA technology, launched services featuring high-performance "H200 GPUs" in November 2024. The "NVIDIA B300" instance of GMO GPU Cloud, which now incorporates the latest NVIDIA Blackwell Ultra architecture, is a GPU designed for building ultra-large-scale AI factories, equipped with 5th generation Tensor Cores supporting "NVFP4" (*3), HBM3e memory (*4) reaching 2.1TB, and high-speed communication via 5th generation NVIDIA NVLink and NVLink Switch. This significantly reduces training time for large language models compared to the previous H200, substantially improving AI development efficiency. Through this service, GMO Internet provides companies and research institutions working in the generative AI field with optimized infrastructure and flexible, customizable computing environments tailored to customer workloads, contributing to shortened development periods and cost reduction while promoting the development of Japan's AI industry.

(*3) "NVFP4": a 4-bit format purpose-built to deliver exceptional inference latency, throughput, and efficiency—all while maintaining production-grade accuracy.

(*4) "HBM3e": The latest high-bandwidth memory standard with improved bandwidth compared to previous versions, accelerating training and inference processing for large-scale AI models.

About GMO Internet, Inc.

GMO Internet, Inc. launched with a new structure on January 1, 2025, to integrate the strengths of GMO Internet Group's Internet Infrastructure business and Advertising & Media business.
Maximizing the solid revenue foundation of the Internet Infrastructure business and the respective strengths of the Internet Advertising & Media business, under our corporate tagline "Internet for Everyone," we deliver "smiles" and "inspiration" to all stakeholders and challenge ourselves to create value that shapes a new future with AI.