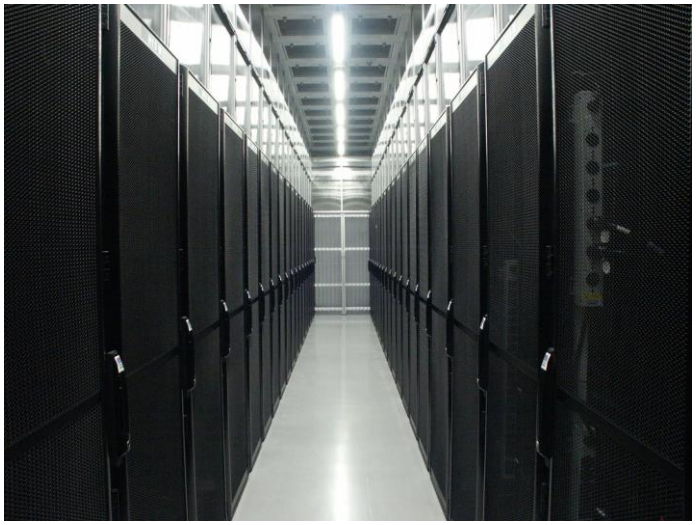
SAKURA INTERNET KOUKARYOKU CLOUD SERVICES ACCELERATE AI SERVICES

Sakura Internet Koukaryoku Creates New Services with Supermicro's 8U SYS-821GE-TNR System, Which Incorporates the NVIDIA HGX™ 8-GPU AI Supercomputing Platform, Accelerating Generative AI Applications



INDUSTRY

Cloud Service Provider

Introduction

Sakura Internet Inc. is Japan's leading cloud service provider with a long history. In addition to its standard server services, Sakura Internet Inc. has adopted the high-performance AI-optimized GPU server Supermicro SYS-821GE-TNHR HGX 8GPU server and launched Koukaryoku cloud service to meet the rapidly growing demand for AI services from training to inference, including generative AI, large-scale language processing, image analysis, and more.

CHALLENGES

GPUs required for generative AI are expensive.

High level of expertise is required for setting up the physical GPU server environment

Sakura Internet Koukaryoku Cloud Service

The Koukaryoku cloud service is Sakura Internet's computing resource service designed to deliver results at the forefront of AI and deep learning by eliminating the risks of operating and capitalizing on computing resources. It targets optimal ultra-advanced GPU servers that meet the needs of those who want a competitive edge while providing high-performance and cost-effectiveness. Sakura Internet Koukaryoku cloud service solves the following GPU challenges:

- Expensive GPUs are required to develop and deploy generative AI. A high level of expertise is required to set up the physical GPU server environment.
- Legal concerns about foreign GPU services.

Features of Sakura Internet Bare Metal Series Koukaryoku PHY Cloud Service

The Koukaryoku PHY Cloud Service is designed to analyze complex datasets and train large models.

- Bare metal delivers high GPU performance

Koukaryoku PHY cloud service is powered by NVIDIA H100 Tensor core GPUs to meet the demands of a wide range of customers. Providing unparalleled GPU performance that is critical for today's AI development, the Supermicro SYS-821GE-TNHR is powered by eight NVIDIA H100 Tensor Core GPUs to deliver high performance in large-scale language models (LLM), generative AI, machine learning, and scientific simulations.

- High performance, minimal environmental impact.

Sakura Internet's Ishikari Data Center takes an advanced approach to renewable energy, achieving zero annual CO2 emissions. This approach has resulted in high-performance GPU servers running while minimizing environmental impact.

- Accelerate innovation in all fields.

Whether training complex models, implementing algorithms, deep learning image analysis, natural language processing, or scientific simulations, Koukaryoku PHY cloud service accelerates research and business innovation in all fields.

- Large Language Models (LLMs) are ideal for training LLMs in data analysis and risk prediction.
    - Generative AI: for generating a wide variety of content, including images, text, and audio.
    - Scientific Simulation: For complex scientific calculations and simulations.
    - Image Analysis: For advanced image analysis, including real-time image processing.

SOLUTION

Supermicro HGX H100 8-GPU Server

SYS-821GE-TNR

- Dual 4th Gen Intel Xeon Platinum 8480+ Processor
- 2TB DDR5-5600 MT/s Memory
- 8x NVIDIA H100 Tensor Core SXM GPU
- 900 GB/s NVIDIA NVLink™, NVSwitch™

GPU server configuration (per node) for the Koukaryoku PHY

GPU server	Supermicro
Product name	SYS-821GE-TNHR
Cooling method	Air-cooling
CPU (Dual Socket)	4th Gen Intel® Xeon® Platinum 8480+ (56 コア)
GPU	NVIDIA HGX™ H100 AI platform NVIDIA H100 Tensor core SXM GPU x 8
Memory	2TB, DDR5-5600MTs
Storage	7.68TB x 4
Local Ethernet lines, Broadband lossless network	400Gbps x 4
GPU Server node	**260 nodes**

Solution

The Koukaryoku PHY is ideal for data analysis, text analysis, image analysis, etc., as GPU servers are available on a monthly basis (minimum usage period is two months), which is ideal for machine learning and deep learning.

- Available for a wide range of customers, from startups to enterprises.
- Provided as a bare-metal GPU server release service with Linux as the primary OS, customers are free to select and use any framework, API, or tool that meets their requirements.
- Consists of 260 GPU server nodes with a total of 2080 GPUs. Each GPU server node is available by dividing the usage range (node) for each user to provide GPU service to more customers.

BENEFITS

- Latest GPUs to Improve Competitiveness of AI Needs for Business
- High Memory Allows for Large AI Models to be Trained

- Interconnected via Ethernet between 260 nodes of GPU servers, providing multi-tenancy and flexible management capabilities that are important for providing this service as a public service. Network bandwidth will be upgraded to 400 Gbps and other higher bandwidth environments as technology and related products become available.
- The 4th Intel® Xeon® Platinum processors and GPUs are compatible with the customer's applications and stable in operation.

Advantages and Benefits

Renting the latest high-performance GPU is a competitive advantage for AI needs, as it requires no capital investment in server operations, air conditioning, electricity, etc., and a wide range of new models are always available. Sakura's existing general-purpose server and these new GPU servers for AI differ in various aspects and specifications, such as processing performance, required power consumption, and wider bandwidth network design. We believe there are unprecedented opportunities in this new market, creating new value for customers. For high-load processing, customers can use the Sakura Internet Bare Metal Series Koukaryoku PHY.

"We have been a long-term user of a wide range of Supermicro's products in large numbers. Supermicro's products offer many advantages to our business, such as rapid response to new technologies, flexibility in hardware configurations, and a wide range of purchasing options. We believe that our adoption of the SYS-821GE-TNHR results from a good fit between Supermicro's strengths and our needs. We will continue to utilize Supermicro's products and contribute to the continuous development of society as a digital infrastructure company." - Takefumi Sudo, Cloud business division, SAKURA Internet Inc.

For More Information:

Sakura Internet Koukaryoku PHY detailed information, please visit here.

Supermicro GPU Servers – https://www.supermicro.com/en/products/gpu

SUPERMICRO

Supermicro is a global leader in high performance, green computing server technology and innovation. We provide our global customers with application-optimized servers and workstations customized with blade, storage, and GPU solutions. Our products offer proven reliability, superior design, and one of the industry's broadest array of product configurations, to fit all computational need.

Visit https://www.supermicro.com

SAKURA INTERNET

Sakura Internet is an internet company founded in 1996. It provides cloud computing services such as "Sakura Cloud," "Sakura Rental Server," and "Sakura VPS" from its own domestic data centers. Based on the company philosophy of "Turning what you want to do into what you can do," the company develops a variety of services to meet customer needs and proposes DX solutions for all fields.

Visit https://www.sakura.ad.jp