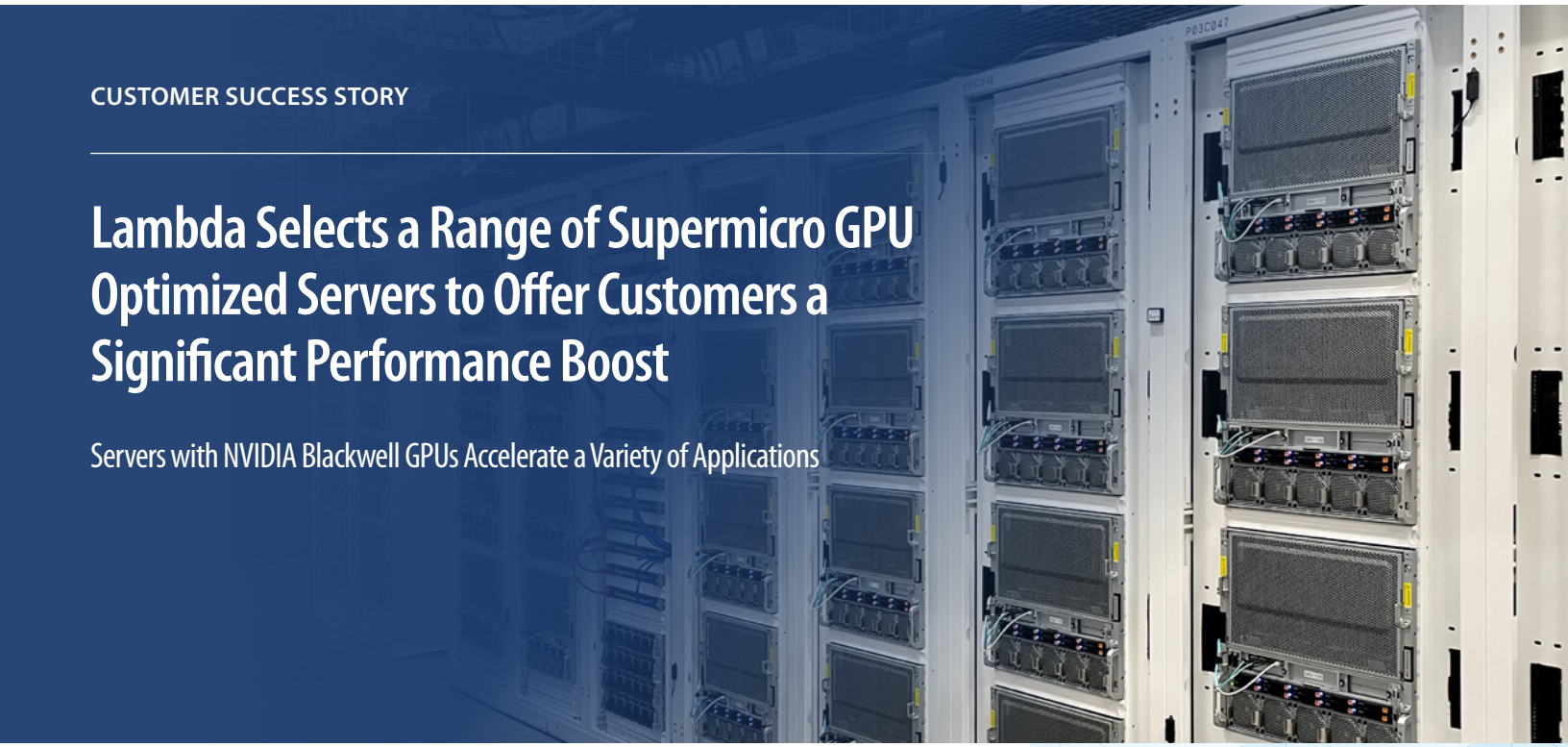
CUSTOMER SUCCESS STORY

Lambda Selects a Range of Supermicro GPU Optimized Servers to Offer Customers a Significant Performance Boost

Servers with NVIDIA Blackwell GPUs Accelerate a Variety of Applications



INTRODUCTION

Lambda is the Superintelligence Cloud. Through Gigawatt-scale AI Factories for training and inference, Lambda provides AI teams with scalable compute to train, and deploy AI at any scale. From rapid prototyping with on-demand compute to serving billions in production, Lambda powers every stage of the AI lifecycle. Trusted by the world's leading AI labs, enterprises, and hyperscalers, they deliver secure, scalable infrastructure purpose-built for AI. With over a decade of experience co-engineering high-stakes cloud solutions, Lambda's teams offer unmatched expertise in AI systems design, deployment, and support at speed, at scale, and without compromise.

Notably, Lambda offers 1-Click Clusters™, giving AI developers instant access to NVIDIA B200 and H200 multi-node GPU clusters, with NVIDIA InfiniBand networking with a simple mouse click.

INDUSTRY

Cloud Service Provider

CHALLENGES

Lambda has been at the forefront of offering customers a choice of systems for their AI development and deployment. As the technology performance and capabilities grew, Lambda has made upgrading its systems to include the latest GPUs from NVIDIA a critical competitive advantage. While the new generation of servers from Supermicro gave them the most performant servers on the market, they also needed to create innovative software that their customers required.

To offer customers the ability to create the desired environment for their AI workloads quickly, Lambda needs a range of servers, which could be used based on the customer's requirements.

SOLUTION

Lambda decided to acquire a significant number of Supermicro servers for its expansion and to serve its customers better. The servers include:

- Intel-based servers with several different CPUs and NVIDIA GPUs

SKU	Processor	Memory	GPU
 **SYS-A21GE-NBRT**	Dual 5th Gen Intel® Xeon® Scalable processors, 8570	3TB DDR5 memory	NVIDIA HGX B200 8-GPU 180GB HBM3
 **SYS-821GE**	Dual 4th Gen Intel® Xeon® Scalable proccessors, 8480+	2TB DDR5 memory	NVIDIA HGX H200 8-GPU 141GB HBM
 **SYS-221HE-TNR**	Dual 4th Gen Intel® Xeon® Scalable processor, 6448Y	2TB DDR5 memory	-

CHALLENGES

- Acquire State-of-the-Art Servers with the Latest GPUs
- Increase Offerings to Customers

PRODUCTS/SOLUTION

SYS-A21GE-NBRT

- 2x 5th Gen Intel® Xeon® Scalable Processors
- NVIDIA B200 HGX

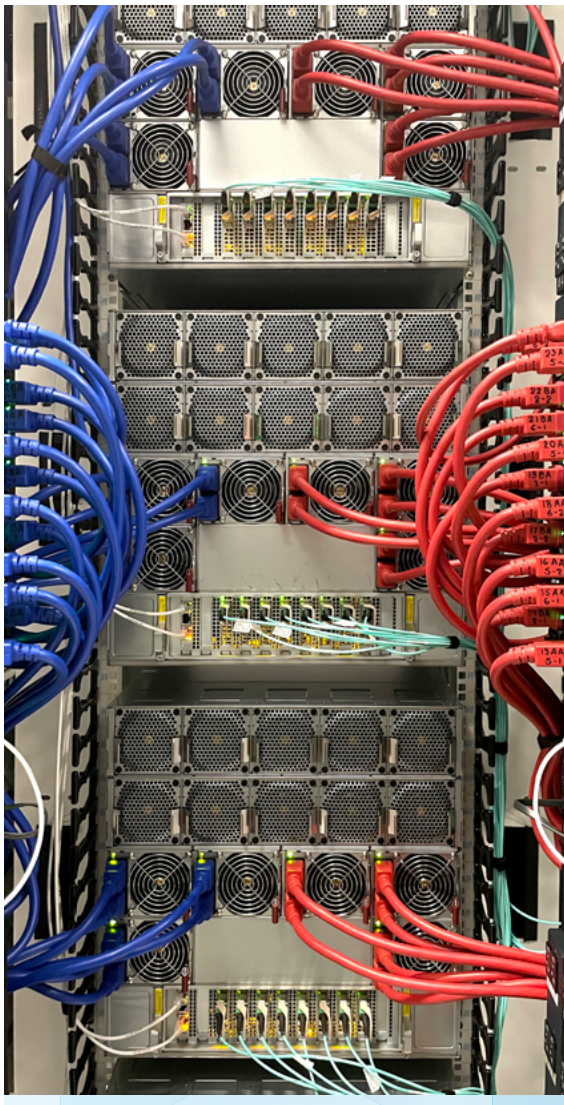
SYS-821GE

- 2x 4th Gen Intel® Xeon® Scalable Processors
- NVIDIA H200 HGX

SYS-221HE-TNR

- 2x 4th Gen Intel® Xeon® Scalable Processors

- Supermicro AI Supercluster, an NVIDIA GB200/GB300 NVL72 rack with:
  - 72x NVIDIA Blackwell B200/B300 GPUs acting as one GPU with a massive pool of HBM3e memory (13.5TB per rack)
  - 9x NVLink Switch, 4 ports per compute tray connecting 72 GPUs to provide 1.8TB/s GPU-to-GPU interconnect
  - Supermicro 250kW capacity in-row coolant distribution unit (with) redundant PSU and dual hot-swap pumps, or 240kW/180kW capacity liquid-to-air solution (no facility water required)
  - Supermicro's end-to-end liquid-cooling solution and comprehensive onsite deployment services
  - Ready for advanced networking technologies for scaling including NVIDIA BlueField®-3 SuperNIC, Spectrum™-X, Quantum-2, and next generation 800 Gb/s networking platforms



The Supermicro GB200/GB300 NVL72 rack is liquid-cooled and is configured as shown below:

Management Networking

- In-band management switch
- Out-of-band management switch

10 Compute Trays

- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Compute Interconnect

- 9x NVLink Switches
- 72 GPUs and 36 CPUs interconnected at 1.8TB/s

8 Compute Trays

- 4x NVIDIA Blackwell GPUs per tray
- 2x NVIDIA Grace CPUs per tray

Liquid-Cooling Options

- Supermicro 250kW capacity coolant distribution unit (CDU) with redundant PSU and dual hot-swap pumps
- 240kW or 180kW capacity Liquid-to-air solution (no facility water required)

> *The depth of Supermicro's server portfolio is an asset to Lambda. As we build Gigawatt-scale AI Factories for Training and Inference, deploy production-grade AI clouds for the world's most AI-advanced organization, and strive to create infinite-scale on-demand compute for our customers, we know we can rely on a partner with a choice of powerful GPU-optimized servers meeting present and future needs."*
>
> – **Kenneth Patchett**
> *VP Data Center Infrastructure, Lambda*

BENEFITS

Lambda can offer a wide range of services to its growing customer base. The availability of a significant number of new servers from Supermicro enables Lambda to provide a wide range of services, which include:

- **1-Click Clusters** – On-demand GPU clusters featuring NVIDIA B200 GPUs with NVIDIA Quantum-2 InfiniBand. No long-term contract required.
- **On-Demand Instances** – Spin up on-demand GPU Instances billed by the hour. NVIDIA H100 and B200 instances at some of the lowest rates in the industry.
- **Private Cloud** – Reserve thousands of NVIDIA H100s, H200s, GB200s, B300s, and GB300s with Quantum-2 InfiniBand Networking.
- **The best-value AI inference** – Access the latest open-source LLMs through a serverless API endpoint with no rate limits, and at a competitive price.

BENEFITS

- Increased offerings for customers
- Faster results from new, GPU-Optimized servers from Supermicro

SUPER MICRO COMPUTER, INC.

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.

www.supermicro.com

LAMBDA

Lambda, The Superintelligence Cloud, builds Gigawatt-scale AI Factories for Training and Inference. Lambda is where AI teams find infinite scale to produce intelligence: from prototyping on on-demand compute to serving billions of users in production, we guide and equip the world's most AI-advanced organizations to securely build and deploy AI products.

Lambda was founded in 2012 by published AI engineers with the vision to enable a world where Superintelligence enhances human progress, by making access to computation as effortless and ubiquitous as electricity.

www.lambda.ai

Printed in USA ♻ Please Recycle 50_2025-Lambda_R07 **AUGUST 2025**