lablup

lablup VULTR

Supercharge Domain-specialized LLMs with NVIDIA HGX B200

Global trade business deploys NVIDIA HGX B200 on Vultr, orchestrates with Backend.AI

Use Case

Supercharge Domain-specialized LLMs with NVIDIA HGX B200

Global trade business deploys NVIDIA HGX B200 on Vultr, orchestrates with Backend.AI

"At Vultr, we deliver high-performance, reliable cloud infrastructure with access to top-tier NVIDIA and AMD GPUs. With 32 global cloud data center regions, we enable businesses to build, deploy, and scale AI applications anywhere—supported by full data residency and compliance, flexible compute options, and transparent pricing. We chose Lablup for its outstanding advancement on workload management and ability to supercharge GPU utilization."

—Kevin Cochrane, CMO at Vultr

Trade needs AI, but infrastructure lags

Global trade operations are highly complex, spanning multiple countries, languages, regulatory environments, unstructured data, and extensive documentation requirements. This complexity increases workloads, while repetitive, error-prone tasks reduce system visibility. The WTO notes that AI integration in trade can significantly cut costs by overcoming language barriers and reducing regulatory and data search expenses. For SMEs, AI solutions can expand access to international trade. (*1)

Deploying trade-specialized LLMs, tailored to each organization's environment and data, boosts competitiveness. As AI rapidly transforms global trade, the gap will widen between companies that invest in infrastructure and those that do not.

Key benefits of AI adoption in trade include:

- Acceleration of trade processes and reduction of operational costs
- Enhanced responsiveness to market changes and emerging risks
- Strengthened compliance in increasingly complex regulatory environments
- Standardization of diverse, non-uniform document formats
- One-stop verification of country-specific regulations

(*1) Source: Trading with intelligence: How AI shapes and is shaped by international trade

Challenges facing tech teams

Many trade companies struggle to realize AI's potential due to the limitations of digital infrastructure and software. Legacy systems, fragmented data silos, and manual, labor-intensive processes remain common. The WTO cites the lack of interoperable infrastructure and standardized processes as key barriers to AI-driven trade innovation.

Even with on-premises AI resources, organizations may lack high-performance GPU servers, and sharing workloads across teams adds complexity. Global deployment requires significant time and expertise. Building a local GPU server is costly, hiring DevOps engineers is difficult, and complying with country-specific privacy laws adds further challenges. Without fast, flexible, and compliant infrastructure—and efficient management software—even advanced GenAI concepts cannot succeed.

Key challenges include:

- Complex and time-consuming infrastructure management
- Insufficient GPU availability when needed
- Inefficient utilization of allocated GPU resources
- Difficulty in building GPU server infrastructure that complies with regional regulations

At a glance

The integration of AI within the trade domain is an essential strategy for the industry. Productivity can be enhanced by introducing AI into human-centric operations. This advancement is realized through the synergy of advanced hardware and intelligent software. The combined solution of Vultr and Backend.AI offers a compelling proposal that addresses these evolving customer needs.

Industry

AI Software

Trade AI across industries

Document automation and classification:

Classify various trade documents and extract key information, improving operational efficiency.

Regulatory compliance review:

Legal information—such as export/import laws, regulations—can automatically review documents, flag potential risks, and reduce human error.

Multilingual translation:

LLMs facilitate translation of trade documents, communication with partners, and interpretation of local laws and regulations, breaking down barriers in global business environments.

About Lablup and Vultr

Lablup develops Backend.AI—Platform that automates infrastructure management with intelligent orchestration and optimizes GPU utilization, enabling GPUs to be allocated and used at the container level. This streamlines the entire workflow —from massive distributed model training to inference, supporting multi-node, multi-tenant deployments—through robust pipeline management. Backend.AI ensures all processes executed securely, dynamically, concurrently in one single software platform.

With instant access to NVIDIA HGX B200 GPUs in 32 global cloud data center regions, Vultr makes it easy to launch high-performance AI infrastructure. Predictable pricing, and built-in compliance support—like ISO 27001, SOC 2 Type II, and GDPR—ensure deployments meet strict requirements.

Challenges

A trading company is developing a domain-specific large language model (LLM) to standardize unstructured documents and comply with country-specific regulations. Fine-tuning with domain-specific data is essential for the LLM to effectively handle various types of unstructured data. However, collecting and curating sufficient data from the actual trade domain requires significant time and human resources. Therefore, training with synthetic data is necessary to ensure model performance and shorten the development timeline. However, generating synthetic data is time -consuming, and all in-house GPU servers are already fully utilized, making it difficult to expand additional workloads.

Solutions

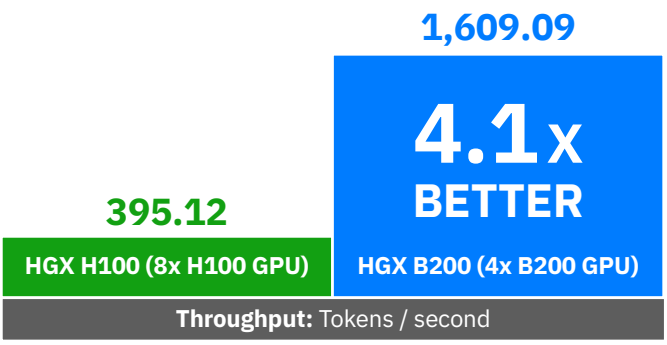
To address these constraints, the company used Vultr Cloud GPUs with NVIDIA HGX B200 acceleration, Backend.AI was deployed for efficient GPU orchestration. The company generated synthetic data using Backend.AI FastTrack 2 MLOps, fine-tuned the LLM, and deployed it. Backend.AI manages workloads across GPUs based on usage, removing the need for user resource management. This ensures efficient, reliable, and scalable model service.

How this benchmark was conducted

Team benchmarked the Qwen-3-235B-A22B model to evaluate performance to build synthetic data. Team selected 100 questions based on 20 examples of dataset, which is 4,000 Q&A sets from 25 trade-related certification exams, Logistics Manager and International Trade Manager. Tests were conducted using vLLM version 0.8.5 with CUDA 12.8 and NVIDIA driver 570. The comparison involved two environments: 4× NVIDIA HGX B200 GPUs on Vultr Cloud GPU and 8× HGX H100 GPUs in the on-prem cluster.

Benchmark: Creating a synthetic data with Vultr + Backend.AI on NVIDIA HGX B200

Lablup and Vultr present compelling solutions for overcoming the technical challenges associated with building trade-specialized large language models (LLMs). By leveraging the combined capabilities of Lablup and Vultr, organizations can establish a fast, secure, and efficient platform for GenAI deployments.



Metrics	HGX H100 (8x H100 GPU)	HGX B200 (4x B200 GPU)
Throughputs Tokens/s	395.12	1609.09
Cost effectiveness	1.3 × unit cost	½ the GPU count
Overall efficiency	-	About 6x Efficient

Results and outcomes

Higher throughput:
By using NVIDIA HGX B200 on Vultr Cloud GPU and Backend.AI, the company achieved 4.1x higher throughput than with in-house H100 GPUs.

Rapid generation of data:
With large-scale batch processing and accelerated computation, data required for LLM fine-tuning could be generated quickly.

Decreased infra setup time:
Cluster setup time reduced to less
than 1/20 of starting from bare-metal, and container-level GPU virtualization maximized GPU use, saving time and costs.

Less resource, greater efficiency:
The B200 matched H100 performance with half the resources, yielding about 6x greater efficiency. Without wasting resources, fully utilize your GPU capacity.

Learn more about Backend.AI and Vultr
Contact us or visit backend.ai to get started.

Additional resources:
Backend.AI - The NVIDIA® DGX™-Ready Software