Solution Brief

Lablup Backend.AI® + Intel® Gaudi® 3 AI accelerator

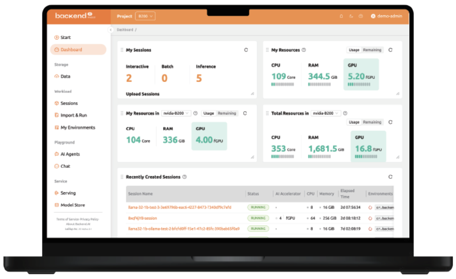
AI at Scale,
Ready for Tomorrow

Lablup Backend.AI® + Intel® Gaudi® 3 AI accelerator

Experience the combination of the maximum performance at scale

Deploy Intel Gaudi 3 AI accelerator with Backend.AI to unlock intelligent orchestration that maximizes your investment. Dynamically allocate resources, enforce multi-tenant isolation, and coordinate heterogeneous hardware through Backend.AI's Sokovan engine. Build AI infrastructure that doesn't just deploy cutting-edge accelerators, but orchestrates them intelligently to balance performance with efficiency.

Backend.AI: Answer for hyper-scale deep-learning & scientific computing



My Sessions: Interactive 2, Batch 0, Inference 5; Upload Sessions

My Resources: CPU 109 Core, RAM 344.5 GiB, GPU 5.20 fGPU

My Resources in nvidia-B200: CPU 104 Core, RAM 336 GiB, GPU 4.00 fGPU

Total Resources in nvidia-B200: CPU 353 Core, RAM 1,681.5 GiB, GPU 16.8 fGPU

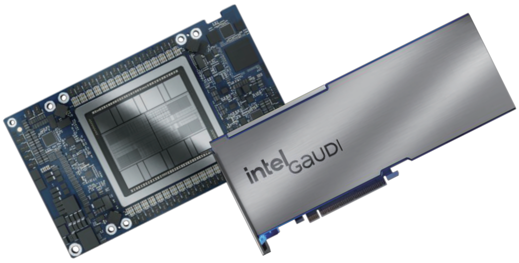
Recently Created Sessions

Session Name	Status	AI Accelerator	CPU	Memory	Elapsed Time	Environments
llama-32-1b-test-3-3e69796b-eac6-4227-8473-7340df9c7efd	RUNNING	-	8	16 GiB	2d 07:56:34	cr.backen
8xqf4JY8-session	RUNNING	4 fGPU	64	256 GiB	2d 08:18:12	cr.backen
llama32-1b-ollama-test-2-bfcfd0ff-15e1-47c2-85fc-390bab65f0a9	RUNNING	-	8	16 GiB	7d 02:08:19	cr.backen

- **Maximize your infrastructure investment with Backend.AI's unified platform that delivers unmatched scalability, reliability, and performance**
- **Join 110+ global sites already managing over 16,000 GPUs with Backend.AI's proven enterprise solution**

Backend.AI is an AI infrastructure operating platform that transforms GPU complexity into operational simplicity. At its core, the Sokovan orchestrator is purpose-built for AI workloads, delivering superior performance and efficiency. Whether you're managing a single GPU or orchestrating massive multi-node datacenters, Backend.AI streamlines management across diverse AI workload types.

Intel Gaudi 3 AI Accelerator: New high-performance option built to handle your AI workloads

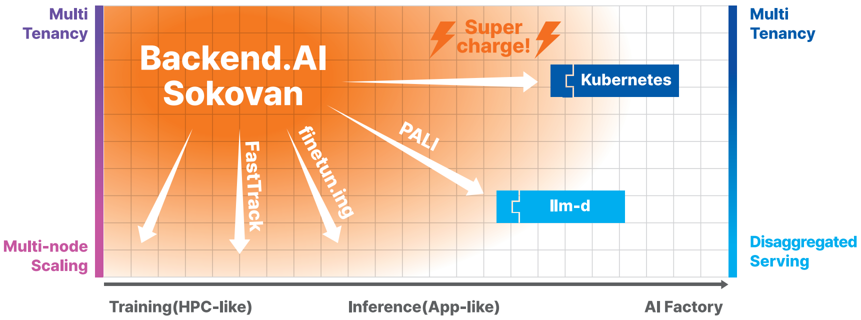


- **Delivering professional-grade performance**
- **Adjusted for demanding GenAI and LLM training and inference workloads**

Intel Gaudi 3 AI accelerator manufactured on the advanced 5nm process, offering a balanced power envelope of around 600W, 128 GB of high-bandwidth HBM2e memory with 3.7 TB/s bandwidth, and 24× 200 Gbps Ethernet ports for high-speed networking and scalable cluster deployment. Available in both OAM and PCIe, the card delivers optimal performance from SMEs to large-scale enterprises.

Maximize your AI ability: Sokovan™ orchestrator

Meet a next-generation AI management system designed for large-scale Intel Gaudi 3 clusters. Flexibly operate without relying on Kubernetes Pod concepts—create containers and allocate resources only when needed. Each session functions as an ephemeral execution environment while supporting persistent storage connections when required, efficiently handling both short-term and long-running workloads.



Dynamic GPU allocation

Backend.AI separates scheduling in two different levels. The first level handles cluster-level node assignment. The second level manages node-level resource and device assignment.

Engineered for multi-tenancy

Isolates user and project contexts through resource groups and scoped configurations rather than dynamic namespaces, improving security and manageability.

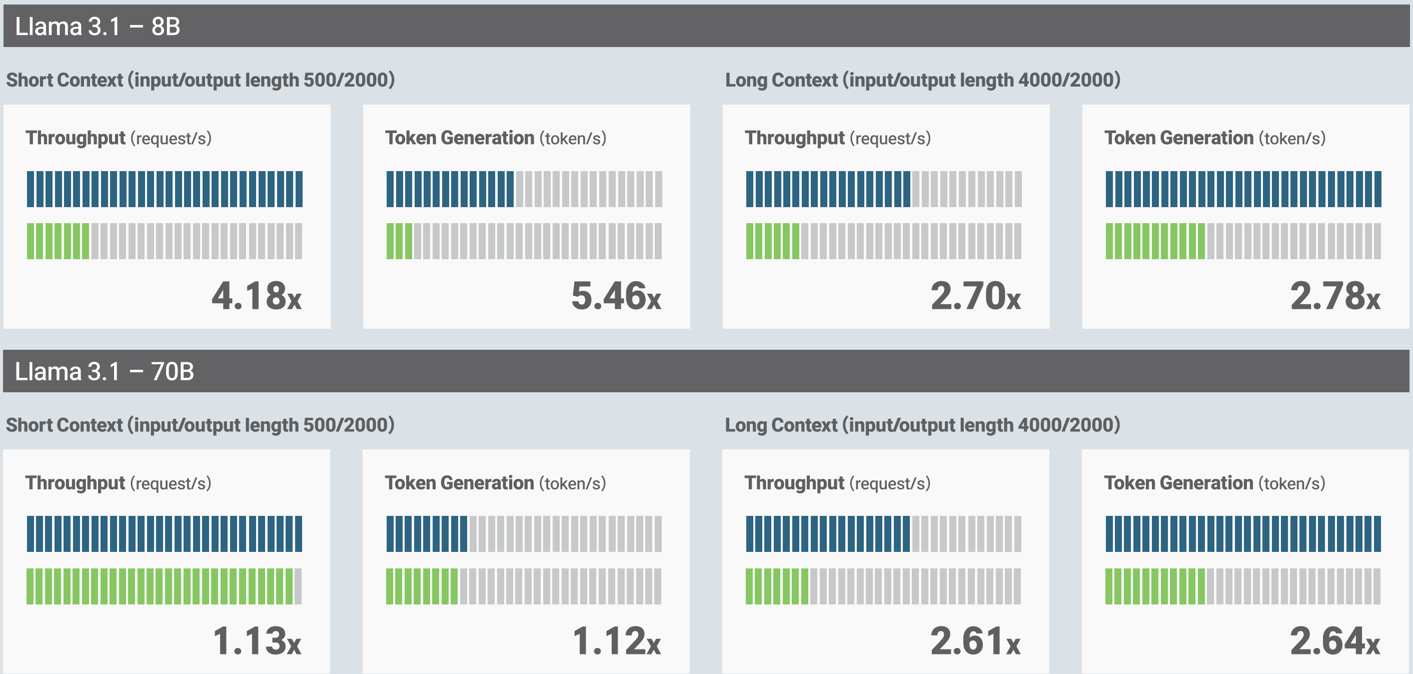
Expanding Kubernetes ability $^{\text{(beta)}}$

Backend.AI's infrastructure management capabilities now extend to Kubernetes. Leverage existing operational platforms while gaining specialized AI workload orchestration capabilities.

Delivering maximum performance of Intel Gaudi with Backend.AI

Tested over Intel Gaudi 3 AI accelerator and Backend.AI 25.14, the Intel Gaudi 3 AI Accelerator consistently outperforms widely deployed accelerators across test configurations, offering throughput improvements ranging between 1.1 and 5.5 times based on model size and context length. It displays particular strength in handling long -context workloads and scales effectively in distributed multi-device setups.

Intel Gaudi 3 AI accelerator vs Comparable Accelerator on Backend.AI 25.14



Disclaimer: Comparable accelerator is the wide-accessible accelerator in the current market situation.
Performance data above is tested on Lablups' internal test under optimized settings. Actual data might vary depending on the environment, or settings.
These benchmarks were conducted on a single Intel Gaudi 3 accelerator unit. The sampled device may not represent the performance characteristics of all Intel Gaudi 3 accelerators in production.
This information is for reference only and subject to change without any notice.

Analysis

Backend.AI delivers Intel Gaudi 3 AI accelerator's maximum performance in every size of context—from edge-scale efficiency to enterprise-grade throughput. In internal benchmark evaluations, Intel Gaudi 3 AI accelerator consistently outperforms competing accelerators on production workloads, addressing both computational intensity and memory bandwidth requirements within a unified architecture.

Small model excellence

Intel Gaudi 3 demonstrates powerful efficiency on compact models. In short-context, it achieves 4.18x higher request throughput and 5.46x faster token generation compared to competing accelerators. Even as context windows extend, Gaudi 3 maintains commanding leads of 2.70x in throughput and 2.78x in token speed, proving its versatility across variable-length workloads and real-time demands.

Large model scalability

For memory-intensive models like Llama-3.1-70B-Instruct deployed with dual-device tensor parallelism, Intel Gaudi 3 shows strong competitive performance. In long-context scenarios, it delivers 2.61x throughput gains and 2.64x faster token generation, validating the architecture's efficiency for extended sequence processing and enterprise-scale inference operations.

Start your AI-ready future

Transform your complex, costly GPU resources into a flexible, high-performance infrastructure that meets the evolving demands. Lablup's robust AI infrastructure OS, Backend.AI, directly addresses intelligent multi -node orchestration, comprehensive multi-tenancy support, and fault-tolerant operation. Intel Gaudi 3 AI accelerators complement this approach with its demonstrated excellence in large language model inference, delivering consistent performance across key metrics.

Learn more about Backend.AI and Intel Gaudi 3 AI Accelerators
bnd.ai/intel-gaudi-3-whitepaper