Solution Brief

AI at Scale, Ready for Tomorrow

Lablup Backend.AI® + Intel® Gaudi® 3

Strategic Integration for NeoCloud Operators

December 2025 V1 rev.0

Lablup Inc.: Jinho Heo, Joongi Kim, Kyujin Cho, Sergey Leksikov, Youngsook Song
Intel Corporation: Arijit Bandyopadhyay, Susan Marquez

Contents

Executive Summary

Backend.AI performance advantages

Backend.AI is a highly efficient AI infrastructure platform that significantly improves large-scale AI inference performance through advanced orchestration and resource management technologies.

Key Strengths:

- Advanced AI workload orchestration up to 4x utilization enhancements
- High-performance with GPUDirect Storage, accelerating tasks up to 5.7x
- Manage every infrastructure at scale, with ultimate efficiency

Ideal scenarios:

- Conducting a multi-node, high-throughput, data-intensive training/inference job
- Operating a multiple, heterogeneous AI accelerator cards
- Multi-tenant, high concurrency deployments

Intel® Gaudi® 3 performance leadership

Intel Gaudi 3 AI accelerator delivers compelling performance advantages, particularly in large language model (LLM) inference workloads, when benchmarked against widely available accelerators in the market.

Key Strengths:

- Consistently superior throughput (1.1–5.5x across workloads)
- Optimized for long-context applications requiring extensive input processing
- Scales efficiently with multi-device configurations

Ideal scenarios:

- High-throughput AI inference (chatbots, content generation, realtime translation)
- Batch processing (summarization, code generation)
- Long-context tasks (legal, research summaries)
- Multi-tenant, high concurrency deployments

Introduction

As AI and deep learning technologies continue to accelerate, enterprises and research institutions increasingly demand higher performance, efficient resource utilization, and scalable operational environments. This technical whitepaper introduces a powerful combination that addresses these evolving requirements: Lablup Backend.AI® running on Intel® Gaudi® 3 AI accelerator systems, especially for the NeoCloud operators.

Backend.AI serves as an AI infrastructure operating platform optimized for Intel Gaudi 3 systems, providing enterprise-essential capabilities including card-level accelerator allocation, NUMA-aware resource scheduling, external storage integration, and user-based storage quota management. This enables researchers and developers to seamlessly operate large model experiments, distributed training, and large-scale inference services without the burden of complex infrastructure management.

The Intel Gaudi 3 AI accelerator built on an innovative 5nm process architecture, delivers optimal performance and cost efficiency for diverse AI workloads ranging from large-scale generative AI and LLMs to traditional classification and voice conversion applications. The Intel Gaudi 3 AI accelerator reliably processes large-scale data and deep neural network computations through its 8 MMEs, 64 TPCs, 128GB HBM2e, and 3.7TB/s HBM bandwidth, while Backend.AI's virtualization and orchestration technologies implement efficient resource isolation of accelerators within clusters.

As Backend.AI serves web-based user interfaces, APIs, and command-line tools, Backend.AI simplifies the development, testing, deployment, and operation of AI workloads, delivering continuously validated stability and convenience to customers across various industries. The platform's advanced scheduling capabilities maximize Intel Gaudi 3 AI accelerator utilization through intelligent workload placement and multi-scale deployment support, from single-card inference to multi-node distributed training clusters.

The combination of Intel Gaudi 3 and Backend.AI represents an industry-leading choice for flexibility, scalability, and adoption of cutting-edge generative AI technologies across both data center and cloud environments. This integrated solution addresses the growing demand for efficient AI infrastructure management while maintaining enterprise-grade security and operational excellence.

Market Overview

Surge in AI Adoption and GPU Demand

The rapid acceleration of AI adoption has triggered an unprecedented demand for GPU resources, fundamentally shifting the priorities of enterprise IT and data center operators. NeoCloud providers have emerged as vital players, enabling modern AI workloads by offering specialized, GPU-focused infrastructure and immediate access to high-performance clusters.

NeoCloud: A perfect solution to adapt GPUs on demand

Globally, demand for GPUs far outpaces available supply, leading to prolonged lead times and development delays over a third of enterprises now face multi-week wait times for GPU access from clouds. Hyperscale public cloud providers captured close to 60% of enterprise AI project allocations in early 2025, but pricing volatility and resource constraints left smaller organizations unable to access necessary compute resources. NeoCloud providers emerged to fill this gap, delivering AI-ready infrastructure tailored for compute-intensive workloads.

A critical challenge for NeoCloud operation

Resource utilization efficiencies
Traditional one-user-per-GPU allocation leads to significant underutilization, especially when high-value GPUs are dedicated to lightweight inference or testing tasks. This directly affects profitability and cost-effectiveness for providers and their customers.

Tenant isolation, security, and compliances
Enterprise NeoCloud demands strict security, rigorous tenant isolation, and compliance with local/global regulations. Managing diverse workloads with robust isolation while ensuring regulatory adherence requires sophisticated orchestration, advanced networking, and continuous certification.

Scaling complexities across infrastructure
Efficient scaling requires orchestrating heterogeneous hardware across multi-node clusters, maintaining high performance and reliability under fluctuating demands. Consistency, failover capabilities, and optimized workload distribution are at the heart of competitive differentiation.

Fault tolerance and High Availability (HA) setups

Ensuring business continuity in the face of hardware failures or maintenance is a non-negotiable requirement for NeoCloud customers. Advanced orchestration and multi-node management allow NeoCloud providers to quickly redistribute workloads and maintain service availability.

Lablup + Intel: Optimal software and hardware at scale

To address the mounting challenges facing NeoCloud providers, Lablup and Intel have forged a close collaboration, combining their expertise to reshape the future of AI infrastructure. By integrating Lablup's advanced AI infrastructure operating platform with Intel's latest Gaudi 3 AI accelerators, the two companies aim to deliver flexible, high-performance, and cost-efficient solutions that enable organizations to overcome the limitations of traditional GPU performance and resource management.

Company Overview

Lablup

Lablup builds products that make it easy to operate and scale AI by addressing real-world infrastructure challenges across research and industry.
Lablup's Backend.AI enables organizations to extract maximum from their infrastructure, combining scalability, reliability, and performance within a unified platform. It supports distributed training, hyperparameter tuning, and large-scale inference with secure, concurrent, and multi-tenant execution. Today, the platform manages over 16,000 GPUs across 110+ sites worldwide—including telecom providers with GPU-as-a-Service operation, financial institutions operating in strictly air-gapped environments to build in-house LLMs with sensitive customer data, manufacturers who integrate AI into their product build environment, and medical organizations analyzing sensitive patient imaging data.

Intel

Intel stands as a global technology leader advancing processor and accelerator innovations at the heart of modern computing. The company is building a comprehensive AI product portfolio spanning multiple market segments. For the AI server domain, Intel Xeon processors, Gaudi accelerators, and its next-generation AI accelerator Crescent Island deliver scalable performance for enterprise workloads. On the consumer side, Intel AI PC platforms and Intel Core series processors maintain market competitiveness across laptops, workstations, desktops, and edge systems. The company excels through its cutting-edge Intel Xeon processors (P-Cores and E-Cores) with in-silicon acceleration engines like Intel AMX, Intel AVX-512 and AVX-2, Technology, QAT, DSA 2.0, IAA 2.0, and more - capabilities, and Intel Gaudi Platform further strengthens its position at the forefront of machine learning acceleration. Built for next-generation AI workloads. Intel empowers organizations to efficiently deploy AI applications across computer vision, natural language processing, and other vertical domains, establishing itself as a premier choice for enterprise AI innovation.

Backend.AI – 'AI Infrastructure OS' of Intel Gaudi 3 AI accelerator

Backend.AI provides a platform that fully leverages Intel Gaudi 3 AI accelerators with advanced orchestration and management in heterogeneous multi-node environments. It efficiently schedules Intel Gaudi 3 resources for optimal utilization, while its support for high-bandwidth networking and deep learning engine which streamlines large-scale training and inference.

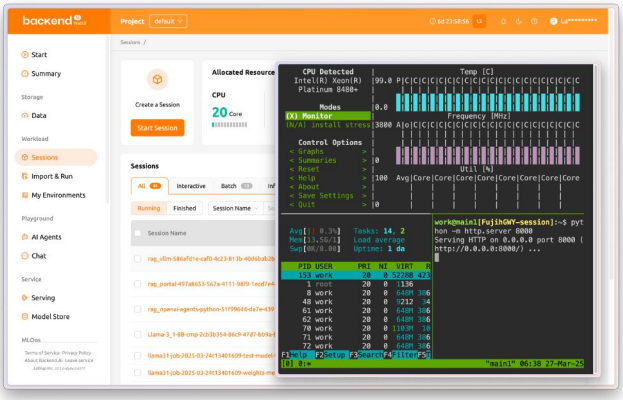


Figure. 1. Backend.AI WebUI

Composable AI Stack of Backend.AI

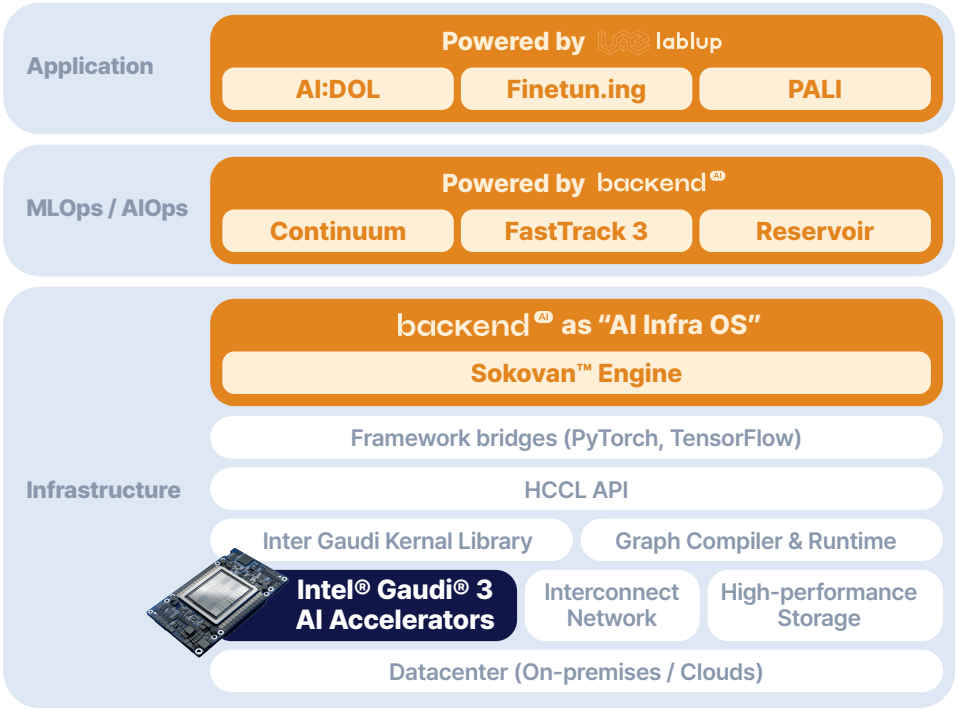


Figure. 2. Complete AI Stack over Intel Gaudi 3 AI Accelerator

Backend.AI delivers a fully composable AI stack that integrates Intel Gaudi 3 AI Accelerators within a unified architecture spanning from infrastructure to application. The Sokovan™ engine orchestrates heterogeneous compute resources—including CPUs and Intel Gaudi 3 accelerators—across multi-node environments, while Backend.AI Core provides comprehensive MLOps capabilities. The ecosystem includes Backend.AI Continuum for intelligent failover in mission-critical deployments, FastTrack 3 for batch-oriented workflow management, and Reservoir AI for model and package management in air-gapped on-premises clusters. This integrated platform enables organizations to develop, train, and deploy AI workloads with optimal scalability and resource utilization.

Sokovan™ – Powering your AI workload on Intel Gaudi 3 AI accelerators

Backend.AI's Sokovan orchestrator, shown in Figure 2, is a next-generation AI cluster manager optimized for large-scale, multi-node environments such as those built with Intel Gaudi 3. Designed specifically for AI workloads, Sokovan operates without traditional Pods, enabling flexible computing sessions composed of on-demand container bundles with no pre-allocated resources. Each session behaves like a transient process with an overlay filesystem, while persistent storage is provided through volume mounts, allowing seamless transitions between ephemeral and durable workloads.

Sokovan's Dynamic GPU Allocation provisions and resizes GPU resources in real time based on workload demands, maximizing cluster throughput and minimizing idle time across heterogeneous workloads. Its cluster-level scheduler supports multiple policies —including heuristic FIFO, DRF (Dominant Resource Fairness), and customizable agent selection—with Head-of-Line blocking avoidance to ensure fair and agile scheduling under heavy multi-user conditions.

Engineered for multi-tenant SaaS environments, Sokovan isolates user and project contexts through resource groups and scoped configurations, improving security and manageability without dynamic namespaces. Integration with identity systems such as SSO plugins or Keystone ensures consistent authentication across shared infrastructure, enabling Backend.AI to deliver scalable, elastic, and securely isolated AI-native computing clusters.

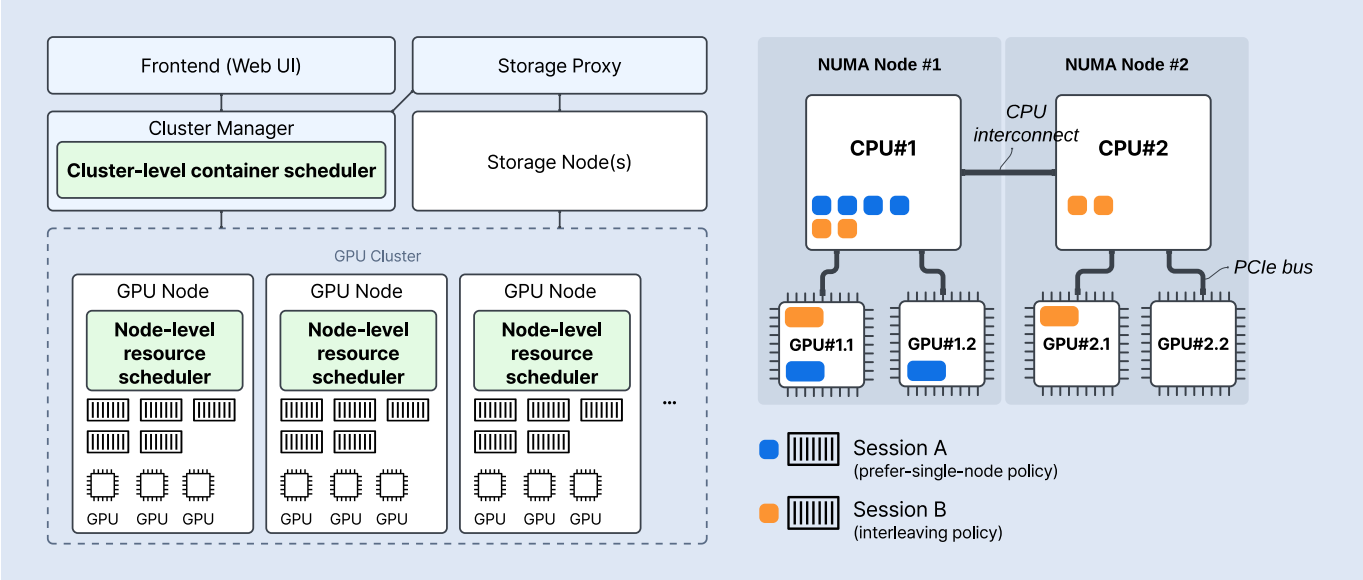


Figure. 3. How Sokovan allocates job requests to the nodes, making Dynamic GPU Allocation possible.

Furthermore, Backend.AI's infrastructure management capabilities now extend to Kubernetes. As part of strategy to enhance market adaptability, Backend.AI adopts a dual-engine architecture that integrates Kubernetes alongside its AI-optimized Sokovan orchestrator. By bridging the operational gaps inherent in standard Kubernetes deployments for AI workloads, Backend.AI enables customers to operate AI-first infrastructure for machine learning and AI applications.

Flexible resource group

Backend.AI enforces robust tenant isolation in multi-tenant Intel Gaudi 3 clusters through comprehensive separation of compute resources, networks, storage tiers, and control planes. Project- and user-level quotas enable precise cost control while maintaining security boundaries, and dynamic rule-based autoscaling continuously adjusts resource provisioning based on real-time workload demands to maximize utilization without compromising performance or data integrity.

The platform's resource group abstraction aggregates hardware into logical units defined by device model, workload type (development, batch, inference), node characteristics, and physical location. These groups can be assigned to specific users, projects, or domains, enabling teams to self-manage workloads and optimize spending within policy constraints. This architecture supports diverse deployment scenarios including exclusive resource assignment for sensitive projects, workload separation by hardware type, and fine-grained control across multi-network or hybrid cloud environments.

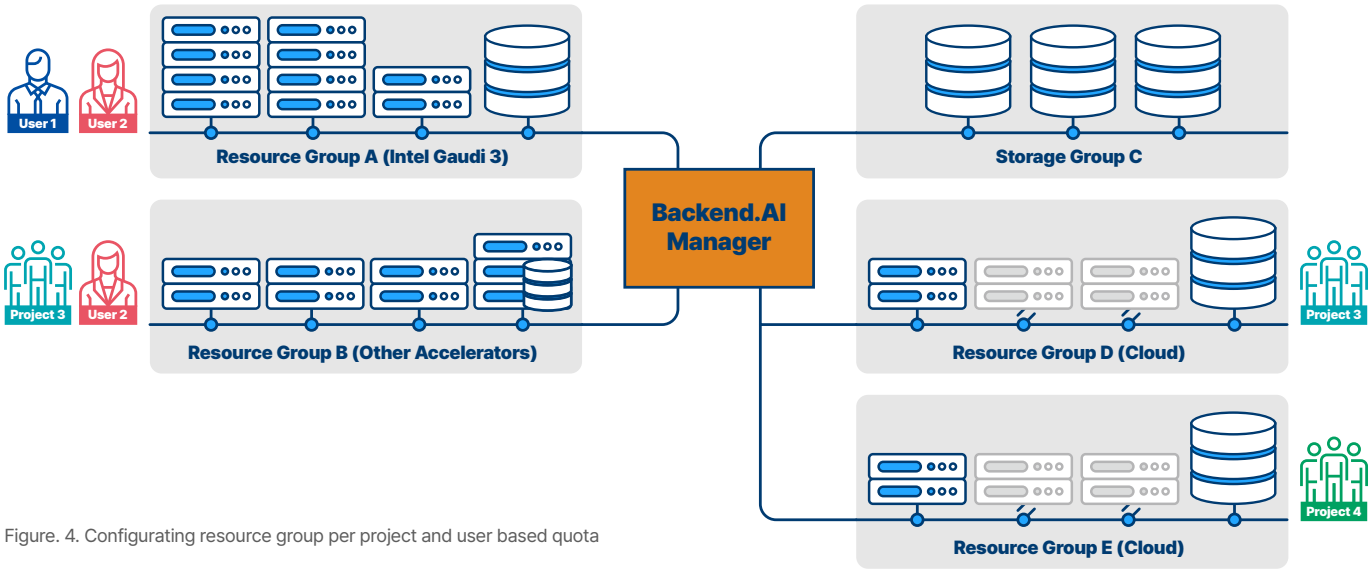


Figure. 4. Configurating resource group per project and user based quota

Backend.AI supports hybrid AI clusters integrating Intel Gaudi 3 AI accelerators along with widely adopted GPUs in the market. This flexibility enables workload distribution strategies that capitalize on the unique strengths of accelerator type, such as assigning different stages to the most suitable hardware. This heterogeneous orchestration enhances performance and reduces cost by maximizing hardware utilization.

Integration with high-performance storage solutions

The platform integrates seamlessly with high-throughput storage systems including Dell PowerScale, VAST, WEKA, NetApp, and more. By supporting a broad range of storage vendors, Backend.AI gives organizations the freedom to work with their existing infrastructure, eliminating compatibility concerns while maximizing the value of current investments. This flexibility ensures rapid data access and efficient distributed cache management, enabling large-scale AI inference on Intel Gaudi 3 clusters with optimal performance and simplicity backed by the storage systems.

Tenant-isolated flexible environment and version management

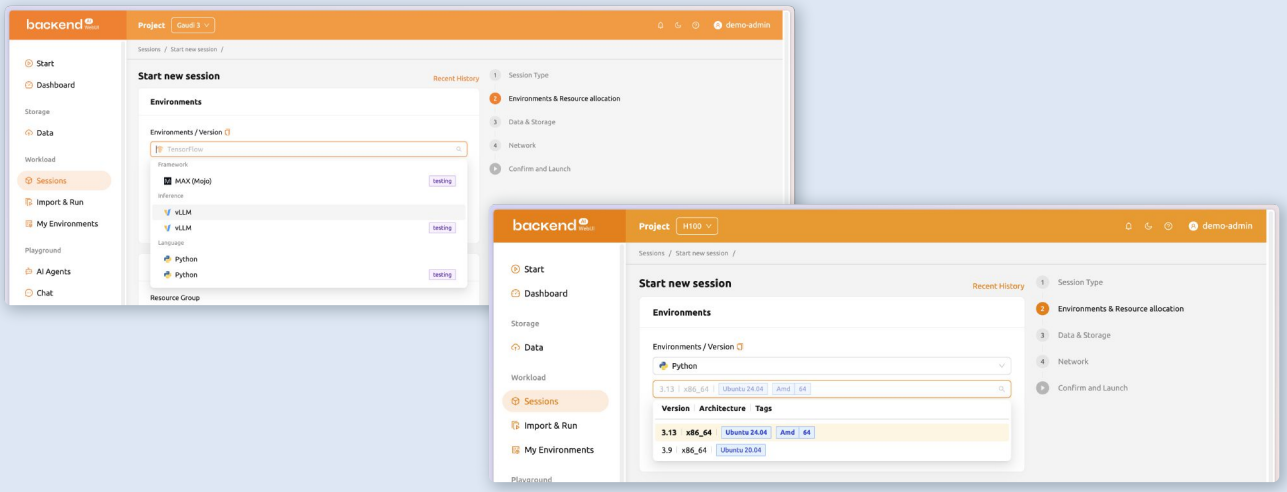


Figure. 5. Backend.AI environment and version selection

No matter the environment or version you choose, Backend.AI keeps everything consistent and under control. The platform allows users to configure their inference, training, and development environments with precise flexibility. Selecting from frameworks such as vLLM for inference, PyTorch or TensorFlow for machine learning and languages like Python or Swift, it also supports multiple version options, enabling users to run legacy or cutting-edge setups as needed. Since teams and researchers often require different configurations for optimal performance, Backend.AI provides tenant-level environment isolation and version management, empowering each user to operate in an environment tailored to their specific workflows without interference or inconsistency.

Enterprise-grade usability and unified management

Backend.AI WebUI is designed to orchestrate and manage every aspect of a large-scale cluster through an easy-to-use graphical interface. Administrators can configure and control all elements of the cluster, including hardware resources, networks, storage, and policies from a single unified interface. WebUI enables real-time monitoring and fine-grained control over all workloads running in the cluster, such as creating, stopping, reallocating, and more.

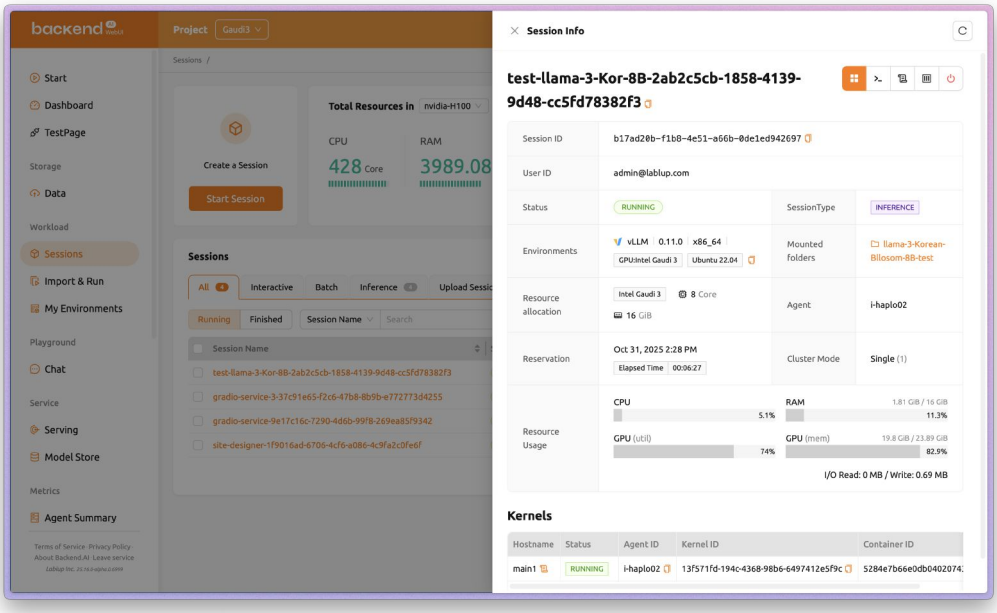


Figure. 6. Backend.AI WebUI

Intel Gaudi 3 AI Accelerator

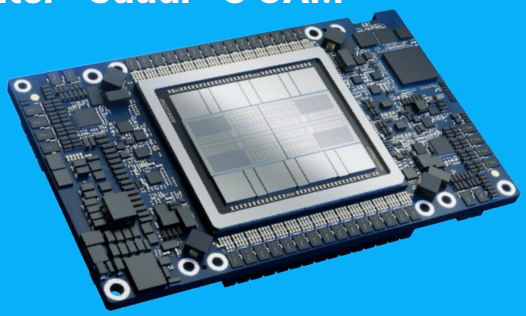
New high-performance option built to handle your AI workloads.

The Intel Gaudi 3 AI accelerator delivers professional-grade performance for the most demanding Generative AI (GenAI) and Large Language Model (LLM) training and inference workloads. Manufactured on the advanced 5nm process, Intel Gaudi 3 offers 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 for both OAM and PCIe version

One architecture can serve two different types of deployment path. From SMEs to large-scale enterprises, Intel Gaudi delivers both OAM and PCIe versions of Intel Gaudi 3 AI accelerator.

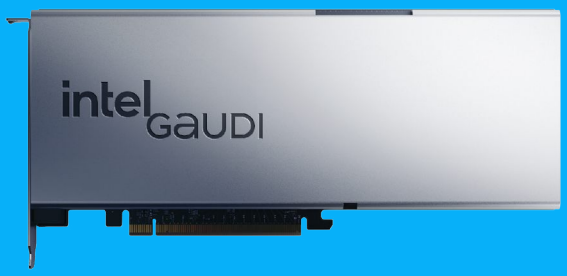
Intel® Gaudi® 3 OAM



Engineered for large scale AI models for hyperscalers, Clusters and Enterprise

- Ideal for running parallel and large LLM inferencing at scale.
- With strong compute power and networking infrastructure it is ideal for AI research labs, hyperscalers, and cloud buildouts
- Designed to handle high-performance, high-bandwidth demands, 6U–8U chassis provide the infrastructure required for full-speed AI acceleration.

Intel® Gaudi® 3 PCIe



Right-sized for Enterprise AI applications and appliances

- Perfect for GenAI inference, secure on-prem workloads, and cost-sensitive AI deployments.
- With its low power consumption and PCIe form factor, Intel Gaudi 3 PCIe is an ideal fit for enterprises, academic institutions, government agencies, and retail branches looking to scale AI without scaling costs.
- Fits seamlessly into 2U–4U servers, enabling more efficient use of rack space and reducing operational overhead.

Feature / Product	**Intel® Gaudi® 3 AI Accelerator**
BF16 MME TFLOPS	1678
FP8 MME TFLOPS	1678
BF16 Vector TFLOPS	28.7
MME Units	8
TPC Units	64
HBM Capacity	128 GB
HBM Bandwidth	3.7 TB/s
On-die SRAM Capacity	96 MB
On-die SRAM Bandwidth (R/W)	12.8/6.4 TB/s
Networking (Bidirectional)	1200 GB/s
Host Interface	PCIe Gen5 X16
Host Interface Peak BW	128 GB/s (64GB/s per direction)
Media Decoders	14

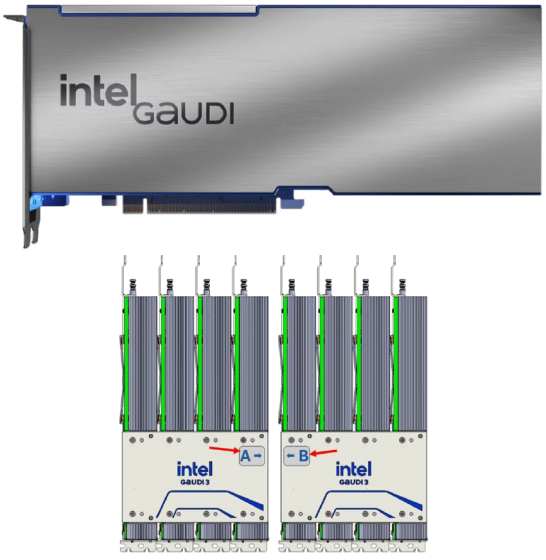
Table. 1. Intel® Gaudi® 3 AI Accelerator Feature

Comm. Type	Datatype	Peak TFLOP/sec
MME (Matrix)	FP8	1678
	BF16	1678
	FP16 (Signed)	459
	TF32	459
	FP32	229
TPC (Vector)	FP8	57.3
	BF16	28.7
	FP16	28.7
	FP32	14.3

Table. 2. Matrix and Vector Compute Capabilities

Intel® Gaudi® 3 PCIe (HL-338)

Architecture	5th Generation Tensor Processor Core
TDP	600W (air cooling)
PCIE	FH 10.5" in length, Double Width (x16 PCIe Gen 5.0)
PCIE Peak BW	128 GB/s bidirectional
Data Types	FP32, BF16, FP16 & FP8 (both E4M3 and E5M2)
HBM	8 x HBM2E
HBM Capacity	128 GB
HBM Peak BW	3.7 TB/s
On-die-SRAM	96 MB
On-die-SRAM BW	19.2 TB/s
System config (1×4)	1 group of x4 via Top Board (HLTB-304A)*
System config (2×4)	2 groups of x4 via Top Boards (HLTB-304A & HLTB-304B)*
Scale-out support	Via Host-NIC Mellanox InfiniBand



*Top board configuration instructions:
HLTB-304A-Customers to use Top Board HLTB-304A to connect one set of 4 PCIe cards per system (one x4).
HLTB-304B-Customers to use the HLTB-304B top board with HLTB-304A to connect an additional 4 PCIe cards, allowing for a total of 8 PCIe cards per system (two x4).

Heterogeneous computing advantage

Intel Gaudi Platform's ASICs represent a strategic shift toward heterogeneous computing architectures that combine specialized processors for optimal AI workload performance. Unlike general-purpose GPUs that are designed for broad computational flexibility, Intel Gaudi ASICs are purpose-built with a heterogeneous design featuring dedicated Matrix Multiplication Engines (MMEs) and programmable Tensor Processing Cores (TPCs) specifically optimized for AI training and inference workloads.

Customers fit with Intel Gaudi 3 AI accelerator

Intel Gaudi 3 AI accelerator is ideal for customers focusing on Cloud, on-premises and Neo-cloud deployments who require support for all Generative AI workloads, including LLM, VLMs, LAMs, Embedding and Re-ranking models for RAG apps and various other Gen AI models for Multi-Agentic Ecosystem Applications encompassing inference, training, Continuous Pre-training, Fine-tuning, RL, Alignments and and deployment of models ranging from over 10 billion to more than one trillion parameters for various requirements.

It accommodates deployment sizes from a single x8 system to large-scale clusters while maintaining a unified fabric based on industry-standard Ethernet. Intel Gaudi 3 AI accelerator supports open-source AI stacks using frameworks such as PyTorch, Megatron, and DeepSpeed, meeting enterprise demands for flexible, scalable, and standards-compliant AI computing infrastructure.

Benchmarks - Remarkable performance of Backend.AI and Gaudi 3 AI Accelerators

Benchmark Summary

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.

Benchmark Conditions

- Tested over Backend.AI 25.14 environment
- Single GPU/HPU test: Llama-3.1-8B-Instruct
- Dual GPU/HPU test: Llama-3.3-70B-Instruct
- vLLM Version: Gaudi 3 (0.7.2), Accelerator H (0.7.3)
- Synapse AI Version: 1.21.0
- CUDA Runtime Version: 12.6
- 64 concurrent requests processed
- BF16 precision applied

Small Model Performance (Llama-3.1-8B-Instruct)

Llama-3.1-8B - Request Throughput Performance

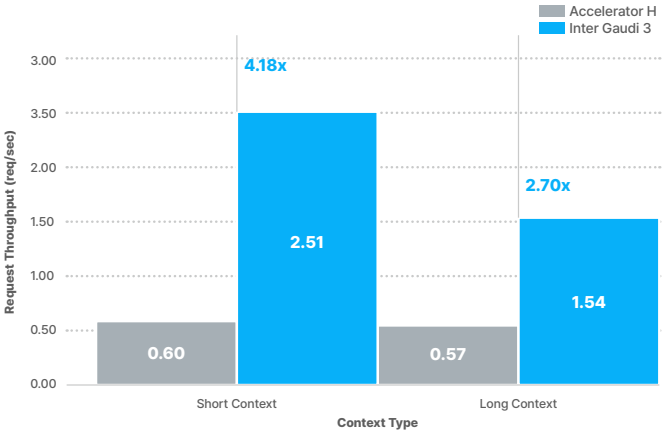


Fig. 6. Llama-3.1-8B - Request Throughput Performance

	Short Context (500/2000)	Long Context (4000/2000)
Intel Gaudi 3	2.51 req/s	1.54 req/s
Accelerator H (Competitor)	0.60 req/s	0.57 req/s
Performance Advantage	4.18x	2.70x

Table 2. Abstract for Llama-3.1-8B-Instruct, Request Throughput Performance

Request throughput performance

The Intel Gaudi 3 AI accelerator demonstrates exceptional request handling capabilities when running the 8 billion parameter model in a single device configuration. In short context scenarios (500/2000 tokens), Intel Gaudi 3 AI accelerator achieves 4.18x higher request throughput compared to widely accepted AI accelerators in the market, processing 2.51 requests per second versus 0.60. For long context workloads (4000/2000 tokens), the performance advantage remains substantial at 2.70x, with Intel Gaudi 3 AI accelerator handling 1.54 requests per second compared to 0.57 for the competitor.

Llama-3.1-8B - Request Throughput Performance

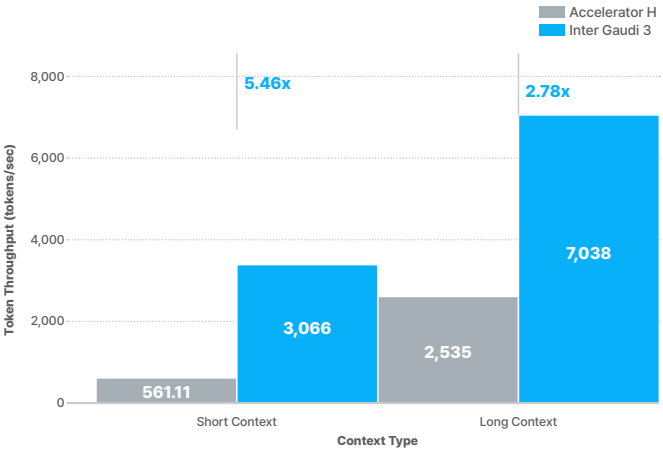


Fig. 6. Llama-3.1-8B - Request Throughput Performance

	Short Context (500/2000)	Long Context (4000/2000)
Intel Gaudi 3	3,066 Token/s	7,038 Token/s
Accelerator H (Competitor)	561 Token/s	2,535 Token/s
Performance Advantage	5.46x	2.78x

Table 3. Abstract for Llama-3.1-8B-Instruct, Token Generation Performance

Token generation performance

Short context configurations show Intel Gaudi 3 AI accelerator achieving 5.46x faster token generation at 3,066 tokens/second versus 561 tokens/second for competitive accelerators. In long context scenarios, the difference remains strong at 2.78x, with Intel Gaudi 3 AI accelerator generating 7,038 tokens /second compared to 2,535 tokens/second. This exceptional token throughput makes Intel Gaudi 3 AI accelerator particularly well-suited for high-volume inference workloads.

Large Model Performance (Llama-3.1-70B-Instruct)

Llama-3.1-70B - Request Throughput Performance

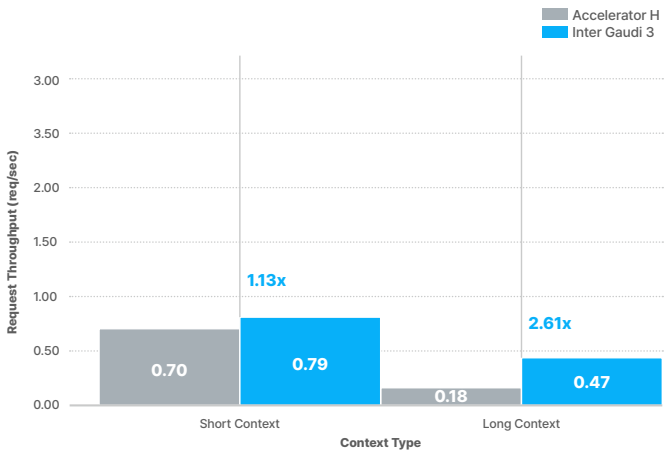


Fig. 8. Llama-3.1-70B - Request Throughput Performance

	Short Context (500/2000)	Long Context (4000/2000)
Intel Gaudi 3	0.79 req/s	0.47 req/s
Accelerator H (Competitor)	0.70 req/s	0.18 req/s
Performance Advantage	1.13x	2.61x

Table 4. Abstract for Llama-3.1-70B-Instruct, Request Throughput Performance

Request throughput performance

When scaling to the 70 billion parameter model using dual-device tensor parallelism, Intel Gaudi 3 AI accelerator maintains competitive performance. Short context workloads show comparable request rates (0.79 vs 0.70 req/sec) with a 1.13x improvement. However, Intel Gaudi 3 AI accelerator excels in long context scenarios, achieving 2.61x higher throughput at 0.47 requests/second versus 0.18, demonstrating superior handling of memory-intensive workloads.

Llama-3.1-70B - Token Generation Performance

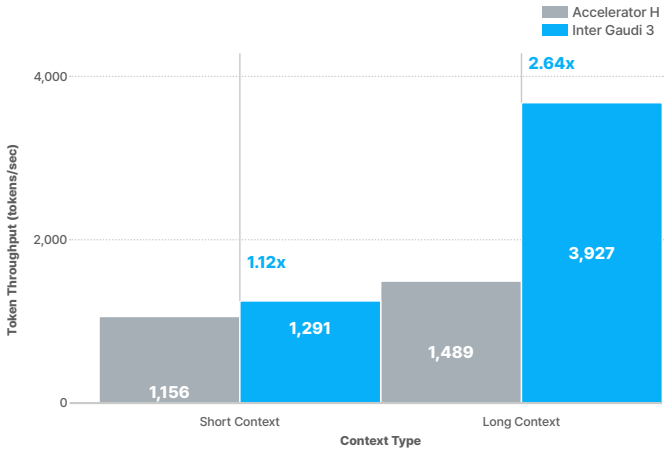


Fig. 9. Llama-3.1-70B - Token Generation Speed

	Short Context (500/2000)	Long Context (4000/2000)
Intel Gaudi 3	1,291 Token/s	3,927 Token/s
Accelerator H (Competitor)	1,156 Token/s	1,489 Token/s
Performance Advantage	1.12x	2.64x

Table 5. Abstract for Llama-3.1-70B-Instruct, Token Generation Performance

Token generation performance

Token generation performance mirrors the request throughput patterns. While short context scenarios show modest improvements of 1.12x (1,291 vs 1,156 tokens/second), long context processing reveals Intel Gaudi 3 AI accelerator's architectural advantages with a 2.64x improvement, generating 3,927 tokens/second compared to 1,489 tokens/second for competitive accelerators.

Key insights

1. Exceptional request throughput for small models:

Intel Gaudi 3 AI accelerator demonstrates outstanding request throughput when running the Llama-3.1-8B-Instruct model on a single device. In short context scenarios, Intel Gaudi 3 AI accelerator achieves 4.18x higher throughput compared to a leading competing accelerator. Even for long context workloads, Intel Gaudi 3 AI accelerator maintains a strong advantage of 2.70x, validating its efficiency in handling varied context lengths.

2. Superior token generation speed on small models:

Token generation performance further highlights Intel Gaudi 3 AI accelerator's strengths, delivering 5.46x faster token speeds than competitors in short contexts. For longer context lengths, it maintains a performance lead of 2.78x, showcasing its suitability for high-volume inference and real-time generative tasks.

3. Competitive scaling performance for large models:

For the Llama-3.1-70B-Instruct model with dual-device tensor parallelism, Intel Gaudi 3 AI accelerator shows competitive request throughput. It slightly outperforms the competitor in short context, and significantly excels in long context scenarios with a 2.61x performance gain, showcasing an architecture optimized for memory-intensive inference workloads.

4. Consistent token generation gains for large models:

Token generation speeds for the 70B model echo throughput trends, with Intel Gaudi 3 AI accelerator achieving a 1.12x advantage in short contexts. The performance gain grows to 2.64x in long context scenarios, underscoring Intel Gaudi 3 AI accelerator’s architectural efficiency for extended sequence processing.

Additional performance appendix

	Precision	Input Length	Output Length	#HPU	Batch Size	Throughput (tokens/s)
LLaMA 3.3 70B	fp8	128	128	8	3986	16622
LLaMA 3.3 70B	fp8	128	2048	8	2048	24705
LLaMA 3.3 70B	fp8	2048	128	8	600	1890
LLaMA 3.3 70B	fp8	2048	2048	8	650	11043

*Data provided by Intel.
**For further information, refer to Intel's official documentation.

Inference Model Performance Data for Intel® Gaudi® 3 AI Accelerators

Deployment recommendations

For smaller models, a single Intel Gaudi 3 AI accelerator is optimal for cost-effective deployment, as it delivers overwhelming absolute throughput performance in this range. This enables significant improvements in concurrent processing and batch efficiency, making Intel Gaudi 3 AI accelerator particularly advantageous for large-scale generation workloads and multi-user concurrent services where total token volume and aggregate throughput are critical.

Models with larger parameter sizes benefit from multi-device tensor parallel configurations. In both the 8B and 70B classes, Intel Gaudi 3 AI accelerator demonstrates a substantial improvement in throughput-per-token (TPOT) when handling long input contexts. This characteristic suggests that deployment strategies should prioritize throughput-oriented batch services.

Optimization strategies

By implementing multiple optimization strategies in combination, Backend.AI deployments on Intel Gaudi 3 AI accelerators can deliver superior efficiency for both throughput-oriented batch inference and multi-tenant service scenarios, while maintaining the flexibility to adapt to diverse workload patterns and performance requirements across enterprise AI applications.

Dynamic batching and Context management

Implement adaptive batching policies to fully experience Intel Gaudi 3 AI accelerator's throughput advantages when serving high request volumes. Context length configuration should be tailored to application requirements, with shorter contexts maximizing tokens-per-second while longer contexts leverage Intel Gaudi 3 AI accelerator’s strong throughput-per-token (TPOT) scaling characteristics.

Multi-tenant orchestration

Leverage Backend.AI's scheduling capabilities to run multiple concurrent inference workloads efficiently, ensuring that Intel Gaudi 3 AI accelerator's high capacity for parallelized requests is fully utilized in multi-user service environments. This approach maximizes resource utilization across diverse workload patterns.

Use of advanced inference frameworks

Deploy modern inference optimization frameworks including vLLM, SGLang, disaggregated inference to enhance scheduling efficiency, memory usage, and token generation performance. These frameworks provide continuous batching capabilities and asynchronous scheduling mechanisms that maintain high GPU utilization even under varying request loads, continuously optimizing with Intel Gaudi platform abilities.

Speculative decoding

Integrate speculative decoding to reduce response latency without compromising quality. This optimization uses a smaller, faster draft model to predict multiple tokens in parallel, which are then verified by the target model, achieving faster inference speeds while maintaining identical output distributions. By leveraging Intel Gaudi 3’s parallel processing and high bandwidth memory architecture, speculative decoding can be executed more efficiently, maximizing utilization and enabling faster token verification and generation cycles for improved overall performance.

Agent aware inferencing

Enhance inference precision and efficiency by leveraging agent-aware inferencing, which adapts computational strategies based on the unique contextual awareness of individual AI agents. This optimization capitalizes on Intel Gaudi 3 AI accelerator’s high parallelism and advanced memory architecture to enable agents to independently perceive, reason, and react within their environments. By integrating Intel Gaudi 3 AI accelerator’s distributed workload management ability, agent-aware inferencing orchestrates complex, multi-agent AI tasks with minimal latency and maximal throughput.

Optimizing based on route LLM

Optimize inference efficiency by dynamically routing queries to the most suitable Large Language Model (LLM) based on query complexity and resource availability. This optimization allocates workloads across multiple LLMs, selecting faster, lower-cost models for simple tasks while reserving high-accuracy models for more complex queries. By leveraging Intel Gaudi 3 AI accelerator's advanced runtime scheduler and heterogeneous hardware support, route-based LLM optimization improves throughput, reduces latency, and balances resource utilization. This flexible routing approach maximizes Intel Gaudi 3 AI accelerator's parallel processing capabilities and memory bandwidth, enabling scalable, cost-effective AI inference across diverse workloads.

Parallelisms and Optimizations

Various parallel processing techniques can be utilized to optimize performance. Tensor parallelism distributes matrix operations across multiple cores, while data parallelism processes input data in parallel. Pipeline parallelism divides model layers into multiple processing stages to maximize throughput. Model parallelism splits large models across multiple devices, and expert parallelism assigns tasks to specialized sub-models within a Mixture-of-Experts (MoE) architecture. Based on Intel Gaudi 3 AI accelerator's high memory bandwidth, scalable network, and multi-core architecture, Backend.AI can effectively fit these parallel techniques to maximize token throughput.

I/O context length optimization

By efficiently managing the context length of input and output data, unnecessary memory and I/O overhead can be reduced, resulting in faster token processing speeds. Combined with hardware optimizations performed at the Backend.AI level, I/O context length optimization allows users to fully maximize the true performance of the Intel Gaudi 3 AI accelerator.

Prefill-decode disaggregation with heterogeneous support

Implementing prefill-decode separation optimizes resource allocation and supports heterogeneous inference strategies. In modern LLM service environments, multiple users' conversations often occur simultaneously, producing requests with varying complexity. When such mixed workloads are processed together, overall request handling can slow down, or complex requests may monopolize system resources, causing performance degradation for other tenants. To address these issues of resource contention and processing-time asymmetry, the compute-intensive prefill phase and the memory-bound decode phase can be scheduled independently, reducing interference and enabling stage-specific optimizations. Since prefill requires higher computational throughput, while decode is more constrained by memory and communication efficiency in accelerator performance, operating these stages separately is an effective way to maximize performance. Customers can handle the prefill phase on their existing hardware while executing the decode phase on Intel Gaudi 3 AI accelerator, leveraging its optimized architecture to deliver exceptional decoding performance. The Intel Gaudi 3's architecture is particularly well-suited for decode operations due to its substantial memory capacity and high-bandwidth memory subsystem. During the decode phase, the model must maintain the entire KV cache in memory while generating tokens sequentially—a process that is memory-bandwidth intensive rather than compute-intensive. Gaudi 3's 128GB HBM capacity per accelerator and 3.7 TB/s memory bandwidth enable efficient storage and rapid access to large KV caches, minimizing memory bottlenecks that typically constrain decode throughput. This memory advantage allows the decoder to serve longer sequences and larger batch sizes simultaneously, maximizing hardware utilization. Backend.AI orchestrates these disaggregated workloads across diverse device configurations, dynamically routing prefill operations to compute-optimized resources and decode operations to memory-optimized Gaudi 3 accelerators, thereby balancing throughput and latency requirements across the inference pipeline.

Conclusion

Lablup's robust AI infrastructure OS, Backend.AI, directly addresses customers' critical needs by providing intelligent multi-node orchestration, comprehensive multi -tenancy support, and fault-tolerant operation. Its orchestration engine dynamically allocates resources across heterogeneous compute clusters, optimizing performance and utilization for diverse AI workloads. Tenants can deploy customized runtime environments that match their specific frameworks, languages, and version requirements. This flexibility allows organizations to maintain compatibility with existing workflows while scaling efficiently across nodes.

Intel Gaudi 3 AI accelerators complement this approach with its demonstrated excellence in large language model (LLM) inference, delivering consistent performance improvements across key metrics. It particularly excels in high-throughput and long-context applications where throughput and latency directly impact customer experience and operational costs. Together, the Backend.AI and Intel Gaudi 3 AI accelerator accelerate NeoCloud operators' ability to offer scalable, reliable, and cost-effective AI services.

Together, these technologies enable NeoCloud operators to transform complex, costly GPU resources into a flexible, high-performance infrastructure that meets the evolving demands of enterprise and research AI workloads.It enables rapid deployment with shorter lead times, support for multi-tenant environments, and optimized cost-performance trade-offs essential for sustaining competitive advantage in the expanding AI economy.

To learn more

AI at Lablup

Backend.AI

Backend.AI on DELL PowerScale

Real-world applications of Backend.AI

Backend.AI with USC Center for Advanced Research Computing

Backend.AI with SKKU Supercomputing Center

Backend.AI with KMU

Backend.AI with Upstage

Backend.AI with KIST

Backend.AI with teamreboott

AI at Intel

Power Your AI Goals with Intel® Artificial Intelligence Solutions

Premier AI Development Resources

Intel Inside – Built for AI

Intel Gaudi Accelerators

Intel® Gaudi® 3 AI Accelerators

Intel® Gaudi® 3 on IBM Cloud

Intel® Gaudi® 3 Whitepaper

Intel® Gaudi® 3 Performance Benchmarks

Intel® Gaudi® Documentation

About Intel Gaudi Software Suite

About Intel Gaudi Management and Monitoring

Dell with Intel® Gaudi® 3

Supermicro with Intel® Gaudi® 3

About Intel® Gaudi® 3 PCIe cards

Intel® Gaudi® 2 on Denvr Cloud

Intel Xeon

Intel® Xeon® 6 Processors

Intel AI PCs

AI PCs Powered by Intel

For more information:

Contant Lablup or Intel for detailed information about this whitepaper.