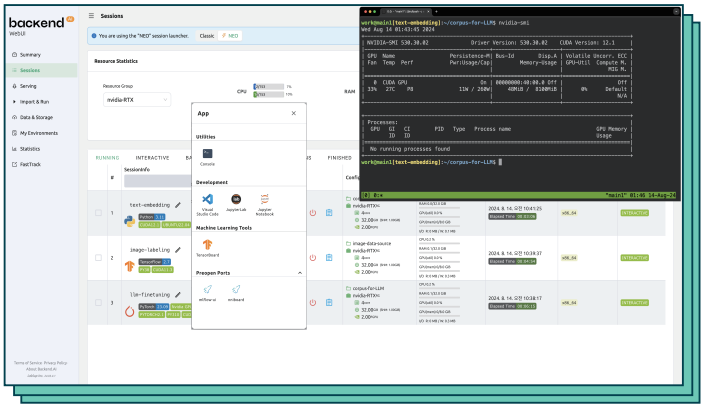
Your business, potential unleashed

Lablup® Backend.AI® meets Intel® Gaudi® 2 & 3 Platform

Get the most out of AI accelerator, Reach your maximum possibilities.

In complex business, promising AI performance and manageability is the key to success. Intel® Gaudi® 3 AI Accelerator, latest release from Intel, offers powerful AI performance and features. Lablup® Backend.AI®, a Platform-as-a-Service, offers a range of features which is optimized for enterprise-level AI environments.



Optimizing Hardware with Advanced Scheduler and Sokovan™

Backend.AI supports a wide range of GPUs and AI accelerators on the market to achieve maximum efficiency of its performance and provides a user interface to make everything easy. This allows customers to efficiently build, train, and deliver AI models from the smallest to the largest language models, significantly reducing the cost and complexity of developing and operating services. Backend.AI is the key to unlock the full potential of Generative AI and Accelerated Computing, transforming your business with cutting-edge technology.

Lablup® Backend.AI® innovations deeply integrated with Intel® Gaudi® 2 AI Accelerators and Intel® Gaudi® 3 AI Accelerators

There are multiple AI Accelerators and GPUs vendors in the market – Backend.AI supports most of them, including NVIDIA, Rebellions, FuriosaAI, AMD, and now, Intel. We worked closely with Intel to provide the best performance in best effort. Finally, we were able to introduce our latest integration with Intel® Gaudi® 2 and Intel® Gaudi® 3 AI Accelerators to the world.

Card-level accelerator allocation

Maximize Intel® Gaudi® 2 and Intel® Gaudi® 3 AI Accelerators cluster workload by lending users the actual number of accelerators intending to. For example, customers can train models on their existing preferred platform, then serve on Intel® Gaudi® 2 & 3 Platform

External storage allocation

Get the most of out the integrated storage solutions in terms of performance. Utilize vendor-specific filesystem acceleration features without user intervention.
(Dell PowerScale, VAST Data, WekaFS, NetApp ...)

Multiple types of workloads

Whatever your environment is, Backend.AI gets you covered. From single-card AI workload which can run small models, to multi-node multi-card AI workload which can run gigantic models, Backend.AI ensure its best performance.

Inference statistics management

Monitor up-to-date, detailed metrics about the performance provided by your AI framework. Backend.AI makes inference statistics management easy, not only showing the information from the hardware, but also on software so that administrator can deep-dive into the metrics.

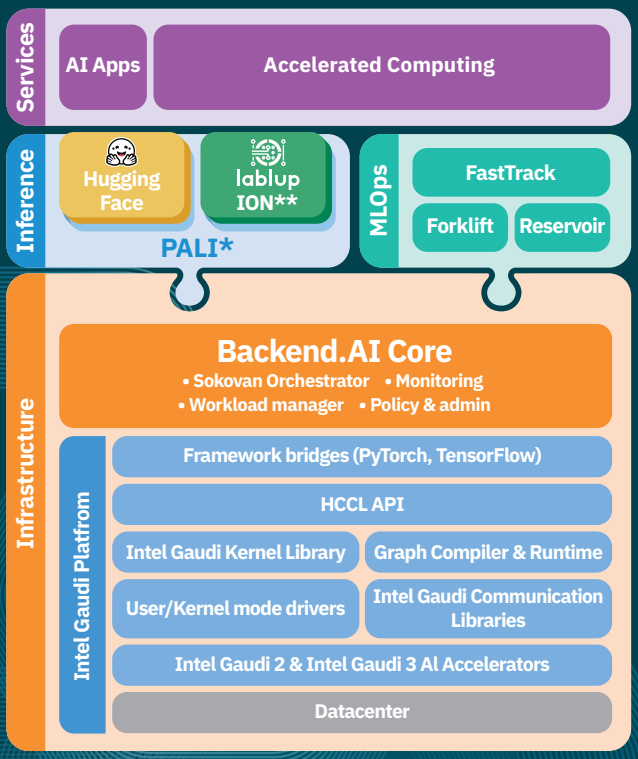
Rule-based inference replica auto scaling

Let system to self-optimize the resource usage. With varied user traffic to inference workloads based on a combination of hardware and software performance metrics, administrators does not need to manually control remaining resources.
* Currently in development (Targeting Dec. 2024)

NUMA-aware resource allocation

Achieve the maximum bare-metal performance by eliminating inter-CPU and PCIe bus overheads within a single node, when there are multiple CPU sockets and multiple accelerators for each socket.

lablup | intel



***PALI** : Performant AI Launcher for Inference ****ION** : Inference Objective Neuralnet

Lablup® Backend.AI®

BareMetal-grade accelerated computing performance

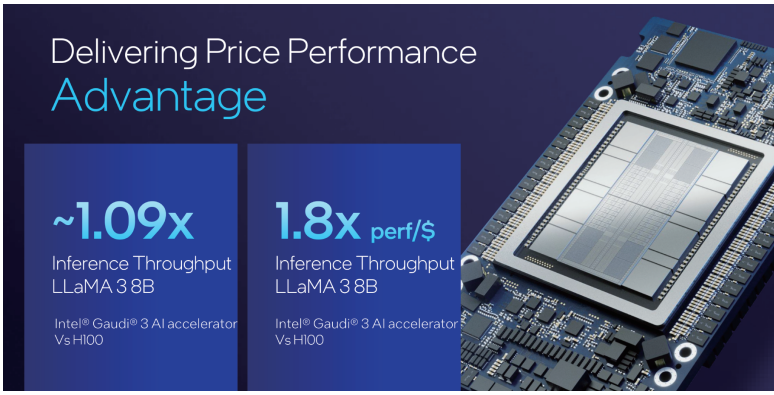
- Outstanding Backend.AI Core performance with our proprietary orchestrator Sokovan™
- Maximizes utilization with idle resource reclamation
- Self-optimizing accelerator resources, combining software and hardware metrics
- Eliminating inter-CPU and PCIe-bus overheads with NUMA-Aware resource allocation
- High-bandwidth, low-latency storage access by RDMA-based storage access provision
- Maximized performance by automated multi-node, high-speed interconnection configs
- Minimizes CPU overhead by Huge Pages allocation

Intel® Gaudi® 3 AI Accelerators

High Performance Acceleration for GenAI and LLMs

Intel® Gaudi® 3 AI Accelerators is driving improved deep learning price-performance and operational efficiency for training and running state-of-the-art models, from the largest language and multi-modal models to more basic computer vision and NLP models. Designed for efficient scalability—whether in the cloud or in your data center, Intel® Gaudi®3 AI Accelerators bring the AI industry the choice it needs—now more than ever.

To learn more about Intel® Gaudi® 3 AI Accelerators, visit **intel.com/gaudi3**



Source

Intel measured results vs H100 data sources: https://github.com/NVIDIA/TensorRTLLM/blob/main/docs/source/performance/perf-overview.md
input-output sequences: 128-2048tps on 1 accelerator/GPU. Intel results obtained on September 9th 2024.
Hardware: One Intel Gaudi 3 AI Accelerators (128 GB HBM) vs one Nvidia H100 GPU (80 GB HBM);.
Software: Intel Gaudi software release 1.18.0.
See Nvidia link for H100 software details Results may vary. Pricing estimates based on publicly available information and Intel internal analysis

Disclaimers

For November 2024, Backend.AI only has support for Intel® Gaudi® 2 AI Accelerators.
Support for Intel® Gaudi® 3 AI Accelerators is planned on the first half of 2025.
Data above is the official numbers provided from Intel, Lablup has no control over these result.
The performance may differ when Intel® Gaudi® 3 AI Accelerators are integrated into the Backend.AI platform.

Make your AI Accelerator "manageable"

We are making AI services efficient, scalable, and accessible to scientists, researchers, DevOps, enterprises, and AI enthusiasts. Lablup and Intel is working closely together to enable the success of Generative AI and deep learning-based services that are popular today.

With our proven technology, Backend.AI provides hardware-level integration with Intel® Gaudi® 2 and 3 Platform for the best effort.

To learn more about Backend.AI®, visit backend.ai

lablup