AI Enterprise

AI Cloud

AI Open Source

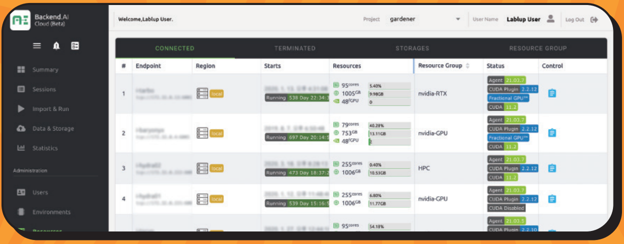
AI MLOps

lablup

backend AI

We'll Get You Every Last Bit.

Backend.AI is a Hyper-Scalable AI R&D platform validated as the first NVIDIA DGX-Ready Software in Asia-Pacific region.



#	Endpoint	Region	Starts	Resources	Resource Group	Status	Control
1	[illegible]	local	[illegible] Running 538 Day 22:34:	95 cores 5.40% 1005 GB 9.98GB 48 fGPU 0	nvidia-RTX	Agent 21.03.7 CUDA Plugin 2.2.12 Fractional GPU™ CUDA 11.2	
2	[illegible]	local	[illegible] Running 697 Day 20:14:	79 cores 40.28% 753 GB 13.11GB 48 fGPU 2	nvidia-GPU	Agent 21.03.7 CUDA Plugin 2.2.12 Fractional GPU™ CUDA 11.2	
3	[illegible]	local	[illegible] Running 473 Day 18:37:	255 cores 0.40% 1006 GB 10.53GB	HPC	Agent 21.03.7 CUDA Plugin 2.2.12 CUDA 11.2	
4	[illegible]	local	[illegible] Running 539 Day 15:16:	255 cores 6.80% 1006 GB 11.77GB	nvidia-GPU	Agent 21.03.7 CUDA Plugin 2.2.12 CUDA Disabled	
				95 cores 54.18%		Agent 21.03.5 CUDA Plugin 2.2.10	

Maximization of GPU utilization

Backend.AI delivers performance and cost optimizations for various deep learning workloads. By harnessing high-performance GPUs connected through a high-speed network, it efficiently handles computationally demanding model training. Moreover, it provides Fractional GPU Virtualization, enabling seamless management of concurrent inference and training workloads. We offer a comprehensive lifecycle plan for your valuable GPU resources, encompassing AI model training, inference services, and machine learning training, guaranteeing the utmost utilization and efficiency.

- **Container-level Fractional GPU Virtualization™ applied**
- **Multi-node distributed GPU training based on via RDMA/Infiniband interconnects**
- **Support for NVIDIA Multi-Instance GPU (MIG)**

Intuitive UI for management

Managing multiple users and tasks in a cluster while fully utilizing all systems can be challenging. Backend.AI provides a simple and consistent user and administration experience, from single node to large multi-node clusters.

- **Web UI/Desktop App**
- **GUI-based MLOps pipeline/batch**
- **Monitoring solution with logs and statistics**
- **Support for CLI/API/SDK for automation and integration**

AI / HPC optimization

Backend.AI utilizes a proprietary GPU-centric orchestrator and scheduler to ensure optimal resource placement and multi-node workloads distribution for AI and high-performance computing. Additionally, it incorporates a storage proxy to parallelize data I/O, further enhancing its efficiency in managing computing resources and unlocking their maximum potential.

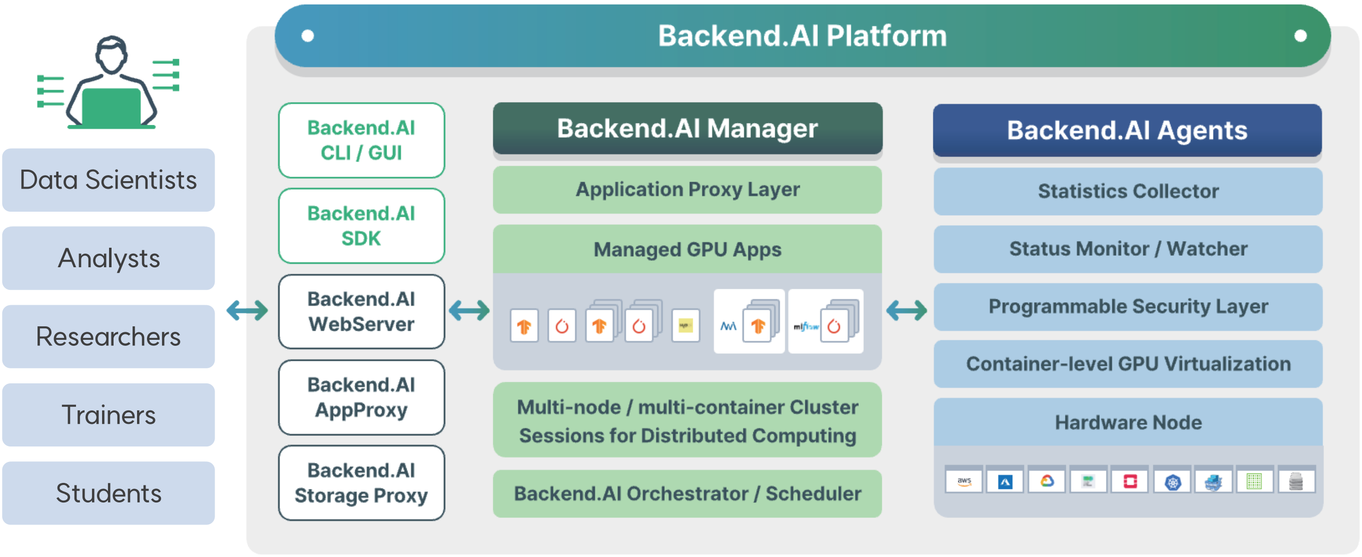
- **Proprietary GPU-centric orchestrator for optimal resource arrangement**
- **Local PyPI/CRAN/APT/Yum repositories for air-gapped clusters**
- **Automatic utilization-based resource reclamation**
- **Batch/pipeline operation scheduling**

Easy scalability of workloads

Begin your deep learning model development with small resources by utilizing Fractional GPU Virtualization. As you progress, seamlessly scale up your operations without encountering any obstacles.

- **Multi-node/multi-container sessions with automatic distributed training setup**
- **Separate model training and data I/O pipelines**
- **Accelerated file I/O support for distributed/ultra-fast storage solutions such as CephFS, PureStorage, NetApp, Dell, Weka.io, etc.**

Backend.AI Platform



Data Scientists

Analysts

Researchers

Trainers

Students

Backend.AI Platform

Backend.AI CLI / GUI

Backend.AI SDK

Backend.AI WebServer

Backend.AI AppProxy

Backend.AI Storage Proxy

Backend.AI Manager

Application Proxy Layer

Managed GPU Apps

Multi-node / multi-container Cluster Sessions for Distributed Computing

Backend.AI Orchestrator / Scheduler

Backend.AI Agents

Statistics Collector

Status Monitor / Watcher

Programmable Security Layer

Container-level GPU Virtualization

Hardware Node

GPU support Container-level Multi & Fractional GPU sharing /NVLink-optimized GPU plugin architecture

Scaling On-premise installation on both bare-metal and VM nodes/Hybrid cloud (on-premise + cloud) and polycloud (multi-cloud federation)/Attaching multiple network planes to containers for data transfers and GPU Direct Connect/GPU Direct Storage in distributed workloads

Scheduling Unified scheduling & monitoring with GUI and CLI admin/Resource allocation per user or user group/Multi-container batch execution and monitoring/Availability-slot based scheduling /Customizable batch job scheduler/Detection and auto-blocking cryptocurrency mining workloads/Automatic policy-based reclamation of idle resources

Security Support for multi-tenancy/Sandboxing via hypervisor or container/Programmable Sandboxing/Syscall-level logging/Administrator monitoring

Reliability High-availability (HA) configuration/On-the-fly addition and removal of compute nodes

UI/UX Desktop application (for Windows 10 & MacOS 10.12, Linux x64 and later)/Web browser support /Control panel and dashboard

Data management Large file transfers via scalable, standalone storage proxy/EFS, NFS, SMB and distributed file system/User & group based access control /Local acceleration cache (SSD, memory)

Developer support Universal programming environments (19 kinds including Python, C/C++, R, Java, MATLAB, etc.)/IDE plugins: VSCode, IntelliJ, PyCharm/ Interactive shell & terminal support/GUI-based custom container image builder

AI developer /data scientist support GUI-based tools (Jupyter, TensorBoard, VS Code, etc.)/NGC (NVIDIA GPU Cloud) platform integration/Fully compatible with major machine learning libraries (TensorFlow, PyTorch, CNTK, Mxnet, etc.)/Concurrent execution of multiple versions of libraries (e.g., TensorFlow 1.X-2.X) /Web-based MLOps with GUI-based pipeline authoring/Automatic update of ML library/DL model as a function/Serving user-written models / Model versioning

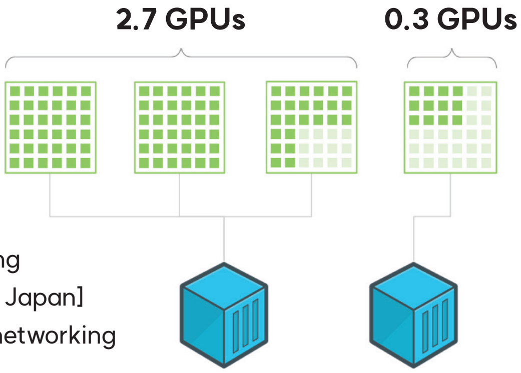
Air-gapped setup Backend.AI Reservoir: a private package repository to serve PyPI, CRAN and Ubuntu/Storage proxy based storage acceleration plugins (CephFS, PureStorage, NetApp, Dell and Weka.io)

Admin System administrator dedicated dashboard/ Administrator dedicated control panel/Compute node system setting control/System statistics /Monitoring solution interlock

Backend.AI Enterprise Package

Essential	Pro	Reservoir
Deep learning research platform for enterprise environments	Total solution and consulting service for service operation and production model development	Package repository integrated with Backend.AI for operation in fully air-gapped environments
Educational institutions and nonprofits	Enterprises, public institutions, and research institutes	Companies and institutions with air-gapped network
ML/AI development environments	Essential included	PyPI/CRAN repository
Web UI and desktop app	AutoML library support	APT (Ubuntu, Debian)
Dev-environment hub	Model service	RPM/Yum (CentOS)
Admin dedicated control panel	MLOps pipeline	Package security check
Fractional GPU Virtualization ™	Model/data store	Package synchronization service
Hybrid cloud configuration	Control dashboard	Dashboard integration

Fractional GPU Virtualization™



Container-Based GPU scaling

- GPU slicing without CUDA MIG and CUDA MPS
- Single GPU sharing: suitable for education and inference workloads
- Multi-GPU allocation: suitable for large workloads such as model training
- Realized with CUDA virtualization layer[Patent registered in Korea, USA, Japan]
- Multi-node distributed GPU training based on high-speed inter-GPU networking

Maximizing AI workload performance and efficiency by integrating the latest high-performance hardware technologies

- Vendor-aware storage I/O acceleration layer (CephFS, PureStorage, NetApp, Dell and Weka.io)
- Accelerated training of large models with RDMA and GPUDirect Storage
- Support for various AI processors to improve performance per watt tailored to workloads
- GPU-NPU integrated pipeline for the maximum performance in minimum cost
- Works with low-power, high-performance computing architecture (ARM & x86)

Backend.AI Reservoir

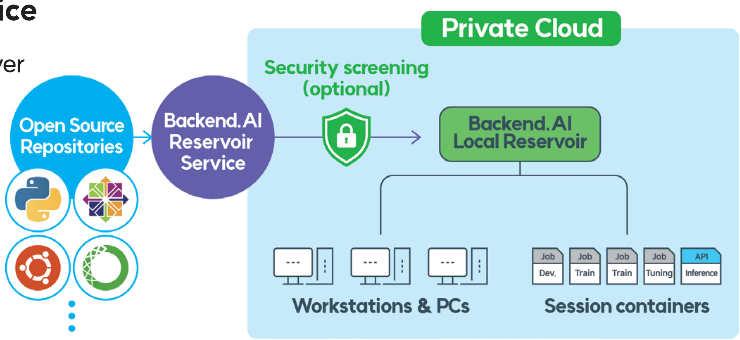
Private package repository for Backend.AI Cluster + Managed update service

Integrated local package repository service

- Lablup Repository Hub Service + Local Package Server
- Provided as an optional Backend.AI Enterprise component

Customer benefits

- Enjoy all-in-one package repositories even in air-gapped setups
- Save the network bandwidth by internal caching



Backend.AI Forklift

Quick and easy GUI tool for container image creation

Streamline the process of building your development environment in just a few clicks, saving your time and effort

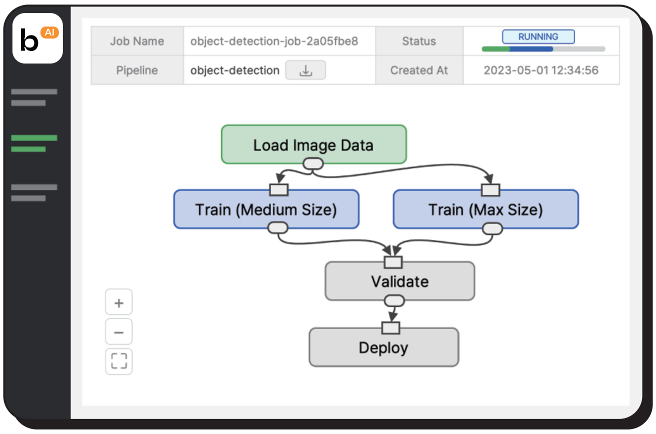
- Quick and easy container image configuration for Backend.AI
- Various pre-installed software packages and default settings
- Easy maintenance and upgrades
- Creates production-ready images for Backend.AI

Backend.AI FastTrack

MLOps platform for maximizing AI engineering efficiency

Manage all aspects of AI development quickly and efficiently

- Provides various AI pipeline templates
- Fully compatible with Backend.AI ensuring optimal performance and cost efficiency
- Provides drag-and-drop GUI for no-code AI pipeline engineering
- Organizes data and code into packages to make it easier to migrate, share, and collaborate
- Secures your AI model service with an auto-configured tunnel proxy



Job Name	object-detection-job-2a05fbe8	Status	RUNNING
Pipeline	object-detection	Created At	2023-05-01 12:34:56

Backend.AI Why should you use?

Backend.AI empowers everyone to rapidly develop and deploy AI solutions on a unified, cost-effective platform. Backend.AI helps you make AI/ML/HPC a part of your business, covering all steps including data analysis, model development, and inference services. Experience the comprehensive capabilities of Backend.AI to make AI a valuable asset for your organization.

Q. I'd like to build a mega farm optimized for running thousands of simulations and analyzing millions of data in a fraction of the time.

A. Backend.AI is tailored for R&D environments such as AI, ML, HPC, and numerical analysis. It optimizes multi-node, large-scale distributed computation by integrating the latest acceleration technologies including GPU Direct RDMA and GPU Direct Storage. It helps analyzing millions of data in a short time by various batch locations, resource allocation, and bottleneck removal specialized for high-performance computing.

Q. While adopting a cloud service for ML education and development, I want to establish and manage the development environment with minimal cost and manpower.

A. Created by ML/HPC experts, Backend.AI offers high availability and increased utilization of GPUs via Fractional GPU™. It does not only reduce costs by providing equivalent performance and educational environments with fewer hardware (GPU) but also help you efficiently manage the system with a small number of people through strong & detailed administration features to handle failures (snapshot-based fail-over, root cause analysis via integrated logging services).

Q. I want to configure Hybrid Cloud by adding Public Cloud instances during periods of high demand.

A. Backend.AI can quickly and easily scale from Public Cloud and on-prem to Hybrid Cloud. In addition, it comes with various GPUs and ML accelerated H/W support and fully documented APIs and SDKs (Python, Node.js) for rapid day-1 configuration.

Q. I want to access my ML development environment anytime, anywhere and only focus on development.

A. Backend.AI offers the convenience of being accessible anytime and anywhere, without requiring any modifications to developer configurations. Simply access it through a web browser, and you're good to go. Additionally, it provides a public API and extensible SDKs, allowing users to seamlessly integrate Backend.AI with their own products, portals, and automation pipelines.

Customer story

Yoon Ho Cho

AI Big Data MBA Vice President, Kookmin University

Backend.AI proved to be highly beneficial for AI education, enabling 80 students to perform modeling exercises and assignments simultaneously using just three GPU servers. Its user-friendly web GUI makes it particularly appealing for educational purposes, as most features are easily accessible. Additionally, Backend.AI offers the capability to configure development environments across multiple servers and efficiently manage resources, even without the need for dedicated administrative staff.

Taegyun Jeon

CEO / Founder of SIA

Backend.AI's GPU virtualization technology made it possible to maximize GPU utilization by eliminating I/O bottlenecks when training large-scale DL models with multi-GPU platforms such as DGX. When we need a lot of resources temporarily, Backend.AI offers on-the-fly addition of Public Cloud instances to configure Hybrid Cloud without the need for complex procedures.

Backend.AI Success story

Backend.AI and NVIDIA DGX are revolutionizing the hyperscale AI infrastructure, running thousands of simulations and analyzing millions of data points rapidly.

Setup

- ML-optimized cluster management solution for multi-organization users in multiple regions
- Hundreds of A100 GPUs and high-performance CPU nodes (for data analysis)
- Completely air-gapped environment for data security powered by "Backend.AI Reservoir," an integrated private package repository

Customer benefits

- Co-design with Lablup for large-scale cluster configurations
- High Availability setup to maximize SLA
- Acceleration of multi-node distributed deep learning via direct inter-GPU networking
- Unrestricted access to PyPI and Ubuntu repositories via Backend.AI Reservoir in air-gapped environments
- Robust system/data security with multi-domain Isolation, safeguarding against concurrent access from both Internal and external users

Representative customers

Companies

SAMSUNG ELECTRONICS kt LG DSME LOTTE CJ

Public & research institutions

한국은행 BANK OF KOREA 건강보험심사평가원 KISTI TTA KAERI KIST

Universities

GIST KMU GNU

Tech & Platform partners

NVIDIA DELL PURESTORAGE NetApp WEKA AMAX KYOCERA KYOCERA Mirai Envision Co., Ltd.

lablup

contact@lablup.com
https://www.backend.ai
https://github.com/lablup/backend.ai