lablup

AI Enterprise

AI Cloud

AI Open Source

AI MLOps

Powering every AI in the world:

backend AI

We'll Get You Every Last Bit.

Seeking methods to enhance your AI services?

Is there an easier way to create or use multiple Generative AIs?

How about training AI models and deploying services?

Massive data I/O causes bottlenecks in AI development. How can we handle this?

How can we manage thousands of GPU clusters and users?

Let Backend.AI be your solution.

Ready-to-use AI Applications

MLOps tools for model training and deployment

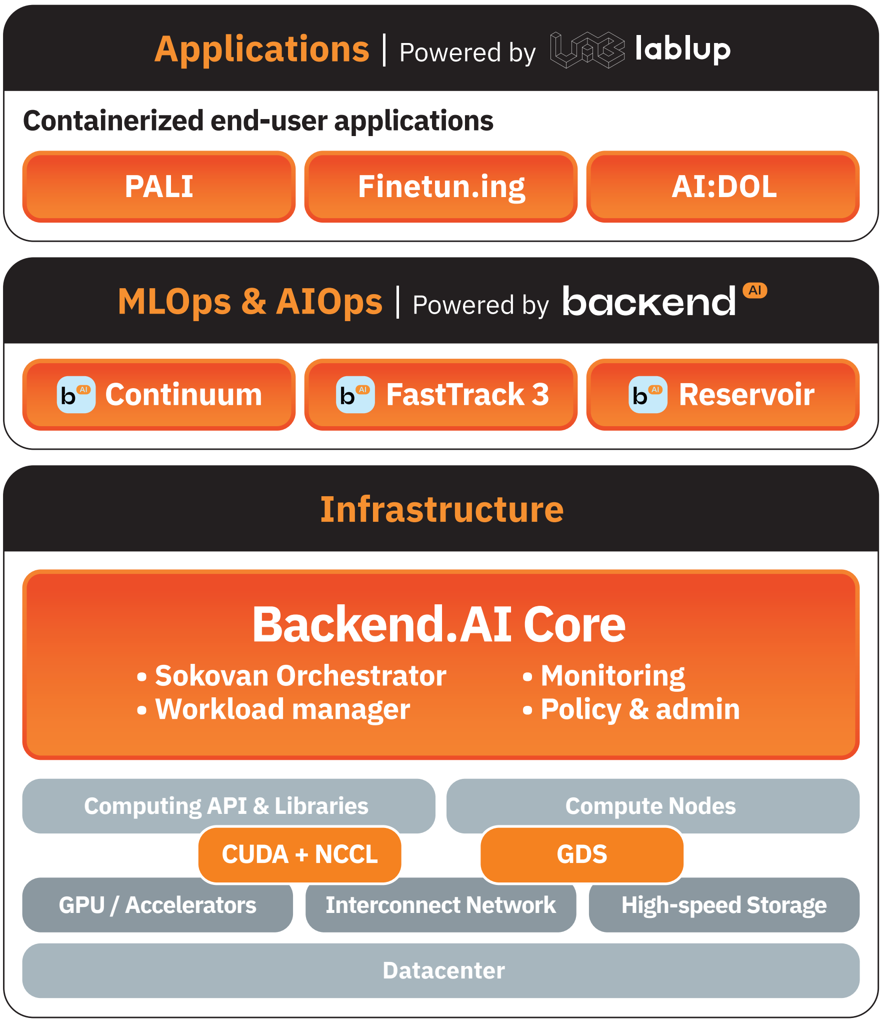
The first solution to implement container-level GPUDirect Storage, in multi-tenant clusters

12,000+ enterprise-scale GPUs in operation

backend AI

Project Map

Discover Backend.AI, built from the ground up for AIs.



Applications | Powered by lablup

Containerized end-user applications

PALI

Finetun.ing

AI:DOL

MLOps & AIOps | Powered by backend AI

Continuum

FastTrack 3

Reservoir

Infrastructure

Backend.AI Core

- Sokovan Orchestrator
- Workload manager
- Monitoring
- Policy & admin

Computing API & Libraries

Compute Nodes

CUDA + NCCL

GDS

GPU / Accelerators

Interconnect Network

High-speed Storage

Datacenter

Connecting GPU management to AI deployment

Ready for your sovereign AI
Supports air-gapped on-premise, cloud-native & hybrid setup

Patented container-level GPU scaling & virtualization
Best-in-class technology powering your GPU efficiency

Sokovan Orchestrator
Intrinsic multi-tenancy & multi-node support

Ready for diverse AI accelerators & multi-architecture
NVIDIA® DGX™ Ready Software and 11+ AI accelerator

Support for data-plane & storage acceleration
Line-rate data plane with RDMA and GPUDirect Storage

Ready for your AI applications

Backend.AI FastTrack 3 | MLOps from model development to service

AI:DOL | Deployable Omnimedia Lab with GenAI

PALI | Performant AI Launcher for Inference

Our technology
Leads the industry

4x

Enhanced Utilization

110%

Pipeline Performance

Our technology
Leads the performance

<3%

Theoretical GPU Overhead

120 Gb/s

GPU Direct Storage Throughput

Our technology
Earns customer's trust

2000+

GPUs on Single site

16k

Enterprise GPUs Managed

Backend.AI Highlights

GPU Supports	Container-level Multi-GPU Allocation / Proprietary Fractional GPU Virtualization / NVLink NVSwitch Optimized Multi-GPU Plugin Architecture
Scaling	On-premises install (Physical/virtual) / Hybrid cloud operation (On-premises + Cloud) and multi-cloud federation / Auto-configures multi-network environments when running distributed workloads
Scheduling	Integrated scheduling, monitoring via GUI Admin (Supports CLI) / User and project-level resource policy / Multi-container batch execution / Slot-availability based scheduling / Expandable, customizable batch scheduler / Detects, blocks crypto mining / Multiple settings to collect idle resources
Security	Supports multi-tenancy / Sandboxing via hypervisor or containers / Additional programmable sandboxing layers / Admin monitoring / SAML 2.0, OIDC, Custom IdP integrations
Reliability	HA (High Availability) setup / Compute nodes hot plugging and scaling
UI/UX	User applications (Windows 10/11, macOS, Linux Desktop) / Web-based service / Control console
Data Access	Upload & download and sharing data via shared storage / EFS, NFS, SMB and distributed file systems / User and project-level access controls / Local cache(SSD/Memory)
Developer Support	Supports 17+ programming languages and runtimes (Python,C/C++, R, Java, etc) / IDE plugins(VS Code, IntelliJ, PyCharm) / Interactive shell, Terminal / Container image builder GUI
Support for AI Devs & Data Scientists	Supports variety of GUI-based development tools in web console: JupyterLab, TensorBoard, VS Code / Integrated NVIDIA GPU Cloud Platform / Supports major ML libraries: TensorFlow, PyTorch, MXNet, Jax, etc. / Ability to use different versions of libraries / Automatic update of ML development environment / Serving serverless deep learning models / Serving and versioning user-authored models
Air-gapped setup	Local package mirror using/by Backend.AI Reservoir(PyPI, CRAN, Ubuntu repository) / Storage-proxy acceleration plugins(PureStorage, NetApp, Dell PowerScale, WEKA, VAST Data, CephFS, LustreFS, GPFS)
Admin & Monitoring	Dashboard and Control Panel for admins / Control compute node & system settings / System statistics collection / Integrate with monitoring solutions / Non- disruptive node management

backend
AI

Have multi-vendor GPUs?
We’ve got you covered.

World's first multi-architecture platform (Arm, x86-64)

NVIDIA intel. GRAPHCORE rebellions_ FURIOSA
tenstorrent AMD groq sambanova HYPER ACCEL

High-performance storage?
We planned for that from the start.

Say goodbye to I/O bottlenecks

DELL Technologies NetApp VAST WEKA
PURESTORAGE ddn IBM Spectrum Scale

Cloud customers,
You’re in the right place.

Experience 95% fast setup compared to bare-metal

aws kt cloud NHN CLOUD
NAVER CLOUD PLATFORM VULTR

100+ customer stories from large enterprises, financial institutions, healthcare organizations, labs, universities, and more.

- NVIDIA ® DGX™-Ready software
- The first solution to implement container-level multi-node GPUDirect Storage

SAMSUNG kt LG HYUNDAI MOBIS 42dot CJ
LIG Nex1 Shinhan Bank THE BANK OF KOREA HIRA HEALTH INSURANCE REVIEW & ASSESSMENT SERVICE
AICA 인공지능산업융합사업단 Artificial Intelligence Industry Cluster Agency ETRI Electronics and Telecommunications Research Institute KAERI TTA KiSTi www.kisti.re.kr Korea Institute of Science and Technology Information
대한민국해군 REPUBLIC OF KOREA NAVY SMC SAMSUNG MEDICAL CENTER GIST Gwangju Institute of Science and Technology
KOOKMIN UNIVERSITY SINCE 1946 KMU 국민대학교 SUNGKYUNKWAN UNIVERSITY 1398 성균관대학교 Korea Institute of Energy Technology KENTECH 2021 한국에너지공과대학교 KOREATECH 한국기술교육대학교 UIPA ULSAN ICT PROMOTION AGENCY

Logos displayed for informational purposes only.

AI Enterprise
AI Cloud
AI Open Source
AI MLOps

lablup

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