Lablup® Backend.AI®

Transforming GPU complexity into operational simplicity

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

Backend.AI enables organizations to extract maximum from their infrastructure, combining scalability, reliability, and performance within a unified platform. Today, 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.

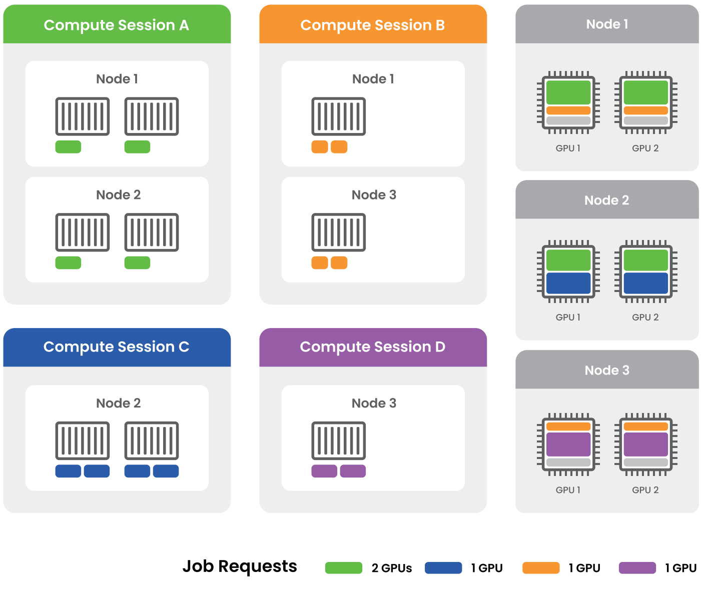
Lablup maximizes GPU utilization to keep pace with the rapidly advancing deep learning software stack, offering the highest level of convenience to users. Backend.AI is NVIDIA DGX™-Ready Software, ensuring seamless operation with NVIDIA DGX systems.

Maximizing GPU utilization for AI/HPC workloads

Optimizing Hardware with Advanced Scheduler and Sokovan™ Orchestrator

Backend.AI Enterprise delivers a high-performance GPU scheduler designed to optimize GPU utilization in large-scale clusters. The platform combines container-level GPU fractioning with granular workload lifecycle management, enabling effective management of dense hardware architectures such as NUMA and RDMA.

Sokovan, Backend.AI's proprietary container orchestrator, addresses the challenges of running resource-intensive batch workloads in containerized environments. Compared to traditional batch scheduling tools like Slurm, Sokovan excels in maximizing the performance of AI workloads by offering acceleration-aware, multi-tenant, batch-oriented job scheduling and seamlessly integrating hardware acceleration technologies. This empowers a wide range of container-based workloads to harness the full potential of the latest hardware advancements. The synergy between Backend.AI's advanced scheduler and Sokovan ensures unrivaled performance and efficiency for both AI and HPC workloads.



How Sokovan™ works

Maximizing Ability of NVIDIA GB200 Grace Blackwell

NVIDIA GB200 NVL72 brings the new computing era, delivering unparalleled performance for LLM inference, retrieval-augmented generation, and data processing. Based on scale-out, single-node NVIDIA MGX™ architecture, its design enables a wide variety of system designs and networking options to seamlessly integrate into existing data center infrastructure.

Backend.AI delivers Software-defined AI infrastructure. Validated on NVIDIA GB200 Grace Blackwell, our platform reliably delivers superior performance at scale. Expand your workload horizontally across NVIDIA GB200 Grace Blackwell while maintaining vertical integration. Eliminate infrastructure bottlenecks with Backend.AI.

Focusing on scalability without compromising usability

Backend.AI is designed to deliver consistent user experience from single-node deployments to large-scale multi-node environments managing tens of thousands to hundreds of thousands of GPUs. Even when operating clusters with thousands of GPUs, the platform maintains system visibility and fine-grained resource control across the entire infrastructure. Built on the design principle of maximizing efficiency without sacrificing flexibility, Lablup optimized Backend.AI to support 11+ heterogeneous GPU architectures available in the market as a unified cluster.

Optimizing workloads with GPU resource management

Backend.AI combines multi-node support, multi-tenant isolation, and patented container-level GPU virtualization to enable GPUs to host diverse size and number of workloads simultaneously. Backend.AI employs dynamic GPU allocation to create and destroy sessions on-demand during workload scheduling. Resources are allocated and reclaimed immediately upon session creation and termination, respectively. This approach gives fine-grained access control over virtualized GPU partitions on GPU nodes without service interruption, enabling on-demand resource provisioning for running services. This approach allows simultaneous execution of a wide range of AI models while ensuring high-performance scaling and robust workload isolation for cloud or on-premises deployments.

Logical resource grouping for precision scheduling and efficiency

Backend.AI's resource management capabilities reflect a level of flexibility and outstanding control that stands apart in the field. Its architecture allows clusters to be partitioned or grouped seamlessly, enabling tailored configurations through logical resource groups. This design supports nuanced scheduling via technologies like agent selectors and job priority settings, facilitating optimized workload balancing within each group. These underlying innovations empower the platform to handle diverse workload types including batch jobs, inference tasks, and interactive sessions with adaptable policies that enhance operational efficiency.

Fast-lane to your containers and storage

Configuring complex firewall rules and keeping up with the network QoS for a mixture of large Backend.AI reduces I/O bottlenecks via dedicated proxy services. These services provide access to your in-container apps (such as Jupyter, SSH, Visual Studio Code) and network-shared storage filesystems, which run separately from the API server. Backend.AI makes ultra-fast storage more accessible to customers by combining acceleration features from ultra-fast storage vendor solutions. Get the most out of the integrated storage solutions, including Dell PowerScale, VAST Data, WekaFS, and NetApp. Backend.AI also supports NVIDIA® Magnum IO GPUDirect® Storage, enabling GPU to directly fetch data from network storage into their memory, bypassing the CPU.

Pre-baked environments and NGC™ integration

Managing ever-evolving software stacks is challenging. Backend.AI simplifies this by offering a variety of pre-built environments, including Lablup's AI software stack, NGC™ catalog images, and full support for NVIDIA NIM™ containers. Our offering covers popular deep learning frameworks, scientific computing tools, and various programming languages, all backed by enterprise-grade compatibility testing. With integrated NIM support, you can seamlessly run, manage, and scale inference microservices, fully automating resource optimization. This ensures immediate access to the latest deep learning and inference environments tailored to your needs.

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

Backend.AI Installation Requirements

Installation requirements subjected to change. Check bnd.ai/requirements for the latest information.

Software	Minimum	Recommended
Operating System	Ubuntu 22.04 RHEL 8	Ubuntu 24.04+ (LTS) RHEL 9+ / Alma Linux 9+
Docker Engine	20.10	25.0+
CUDA	11.0	12.8
PostgreSQL	12.0	16.0
Redis	6.2	7.2

*Open-source version driver recommended

Node type	CPU**	RAM**	Disk**
Manager	2 - 16	2 GiB – 16 GiB	100 GB – 500 GB
Agent	1	512 MiB – 1 GiB	20 GB – 2 TB
App Proxy, Web Server	1 - 8	2 GiB – 16 GiB	10 GB
Storage Proxy	1 - 8	4 GiB – 16 GiB	10 GB (excl. storage volume)
Container Registry	1 - 4	2 GiB – 8 GiB	500 GB – 10 TB

** The ranged values represent the minimum and recommended capacity.