How Upstage, Lablup, and its consortium are powering a national frontier AI model with Backend.AI

Keeping large-scale training resilient:
An end-to-end approach

lablup upstage

Case Study upstage

How Upstage, Lablup, and its consortium are powering a national frontier AI model with Backend.AI

"We spent our time on development—not infrastructure—and Backend.AI handled the rest."

UPSTAGE
Executive Director
Kyle Yi

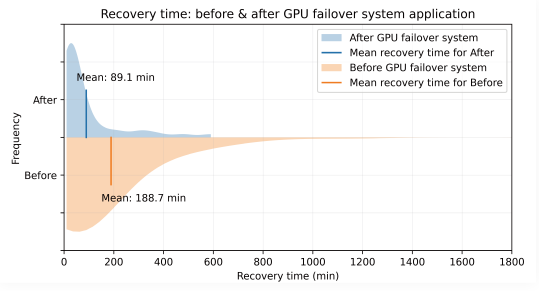
The 'Sovereign AI Foundation Model Project' is a government-led initiative of the Republic of Korea to develop competitive AI models tailored to the Korean ethos. Partnering with Upstage, one of the leading builders of large language models, Lablup joined a startup-driven consortium that passed a highly competitive selection process and was chosen as one of the top five consortia to execute the project. The selected consortia receive support in GPU resources, data, and talent to build AI models that meet global performance standards. This project represents a strategic effort to strengthen Korea's AI capabilities, secure data sovereignty, and promote sustainable growth in the AI sector.

Maintaining continuous GPU cluster performance

The project mandated the consortium to develop a large-scale language model, allowing no room for inefficiency or downtime. To meet this challenge, our infrastructure had to maintain continuous, stable performance under intensive, multi-node workloads. While distributing jobs was straightforward with Backend.AI's mature container-level GPU virtualization and its workload-centric scheduler Sokovan, fleet-scale reality remained: GPU failures are statistically inevitable and hard to predict (see Cui et al., arXiv:2503.11901). To address this challenge, we engineered an end-to-end resilience loop: real-time fault detection through both the widely adopted NVIDIA DCGM and Lablup's proprietary GPU and server monitoring software, automatic alerting to users and operators (MS Teams/ Slack),and—when a workload-stopping fault is detected—automatic node quarantine, hot replacement from a spare pool, and hands-free job recovery that re-launches the multi-node session and resumes from the latest checkpoint. This design minimizes human-in-the-loop effort, sustains throughput on long runs, and preserves researcher productivity. Minimizing human-in-the-loop processes became a key design requirement.

Keeping model training in motion: Backend.AI's automated fault recovery for large-scale training

Upstage leveraged Backend.AI to maintain stable operations across a large-scale cluster of 500+ GPUs. Backend.AI orchestrated hundreds of distributed GPU resources systematically while preserving transparent visibility and control over the entire training environment. This enabled Upstage to reduce training resumption time by nearly 47%, minimizing the effort required for infrastructure maintenance. The team could dedicate their full development time to train models, not managing infrastructure.



Failover time comparison: before vs. after