A need for advanced GPU resource management at the University of Southern California Center for Advanced Research Computing

Solving campus-wide
HPC infrastructure challenges
with Backend.AI

lablup

USC | Center for Advanced Research Computing
Enabling scientific breakthroughs at scale

A need for advanced GPU resource management at the University of Southern California

"If students cannot truly make use of the resources their school provides, then those resources do not truly serve the students. Backend.AI is an AI infrastructure operating platform that makes institutional resources both manageable and accessible."

USC
Associate Chief Research Information Officer
BD Kim

The University of Southern California (USC) is known for its strong emphasis on computer science and artificial intelligence, consistently ranking among the top U.S. institutions in the fields. The university fosters collaboration with industry, enabling practical applications and innovation in computing technology.

Navigating the challenge of instant GPU access for students

The University of Southern California (USC) experienced high demand for sharing CPU-based computational infrastructure at the departmental level, where resources could be allocated immediately upon request. As the university plans to acquire new GPUs, there is a strong interest in extending flexible resource allocation technology to GPU infrastructure. Given that deploying GPUs at departmental scale typically involves lengthy procurement and deployment processes, USC seeks a solution that can pool GPU resources spread across multiple units and allocate them as a controlled resource. The need expands to the connection of a unified billing system that can manage usage details. The absence of comparable implementations in U.S. research institutions reflects the substantial technical hurdles that must be overcome to achieve this level of integrated resource management.

Solving campus-wide HPC infrastructure challenges with Backend.AI

Lablup, headquartered in Korea, has been helping organizations and institutions across the globe maximize their GPU infrastructure utilization. Lablup's reputation in the field motivated USC to pursue Backend.AI as a potential path forward. By using Backend.AI's container-level GPU virtualization, dynamic allocation of computing resources based on real-time demand across labs, groups, and courses is enabled. Its high-performance scheduler and orchestrator maximizes GPU utilization, while fully isolated storage supports complex and varied academic requirements. This allowed the university to optimize its limited infrastructure, ensuring efficient use of hardware. Additionally, integration with a billing system aligned with institutional policies benefits students, faculty, and administration by enabling fair resource usage management.