Nebius + WEKA

Accelerate Your AI Pipeline and Maximize ROI

Challenges

- Operational Tax: Manual cloud setup and "integration debt" delay model training by weeks.
- The Outcome Gap: Sub-optimal cloud performance starves GPUs, creating inconsistent inference and lag.
- Wasted Spend: Inefficient data management and idle GPU hours lead to "bill shock" and poor ROI.

Solution

Nebius's industry-leading AI Cloud with WEKA's NeuralMesh AI-native storage to accelerate development, maximize GPU utilization, and scale with enterprise-grade compliance.

Benefits

- 41x faster response times and 90%+ GPU utilization for training and inference.
- Move from sign-up to training in minutes, reclaiming expensive engineering cycles.
- 161% improvement in ROIC with built-in EU sovereign compliance.

Organizations today face a widening gap between AI investment and AI outcomes. While hyperscalers offer commoditized GPU hours, their generic architectures introduce data bottlenecks and "noisy neighbors" that starve GPUs and slow model progress. Nebius and WEKA restore economic efficiency to AI infrastructure by fusing a world-class, purpose-built AI cloud with the industry's only data platform designed for exascale workloads.

Start Faster: From Concept to Production in Record Time

In the race to market, speed is everything. Our complete pipeline solution eliminates the "first mile" friction that stalls innovation—from initial tuning to global deployment. Unlike piecemeal approaches, our environment treats high-performance resources as native attributes. WEKA NeuralMesh is managed as a first-party Nebius service, ensuring your namespace is automatically mounted the moment a node boots. By removing the burden of manual systems integration and "Day 0" plumbing, we dramatically shorten your path to value, allowing your team to focus on models rather than hardware.

The Result: Collapse development cycles from weeks to hours and move to model training instantly.

Run Better: Maximum Reliability for Training and Inference

For modern AI services, reliability at scale is the product. Our solution is engineered to provide the high-bandwidth path required for massive training runs and the ultra-low latency needed for production-grade inference. We specifically address the instability of generic clouds by providing dedicated GPU infrastructure. By utilizing a liquid-cooled environment and a direct data path that bypasses CPU bottlenecks, we drive GPU utilization from industry averages to over 90%. This ensures your training cycles are shorter, your response times are faster, and your most expensive assets are never idle.

The Result: Slash Time-to-First-Token by up to 41x and achieve peak Model FLOPS Utilization (MFU) across the lifecycle.

Scale Seamlessly: Enterprise-Grade Confidence and Compliance

Scaling isn't just about adding GPUs; it's about having the professional expertise and regulatory confidence to grow without risk. Our solution provides on-demand access to a dedicated, at-scale AI infrastructure backed by "white-glove" professional services. You gain the ability to scale massively on the latest-generation NVIDIA GPUs within a framework that meets the strictest European data residency and compliance requirements. Because absolute data sovereignty is built-in, you can scale confidently, knowing your foundation is secure, sustainable, and ready for the next generation of enterprise AI.

The Result: Achieve a 161% improvement in ROIC while lowering your carbon footprint with 100% green-powered infrastructure.