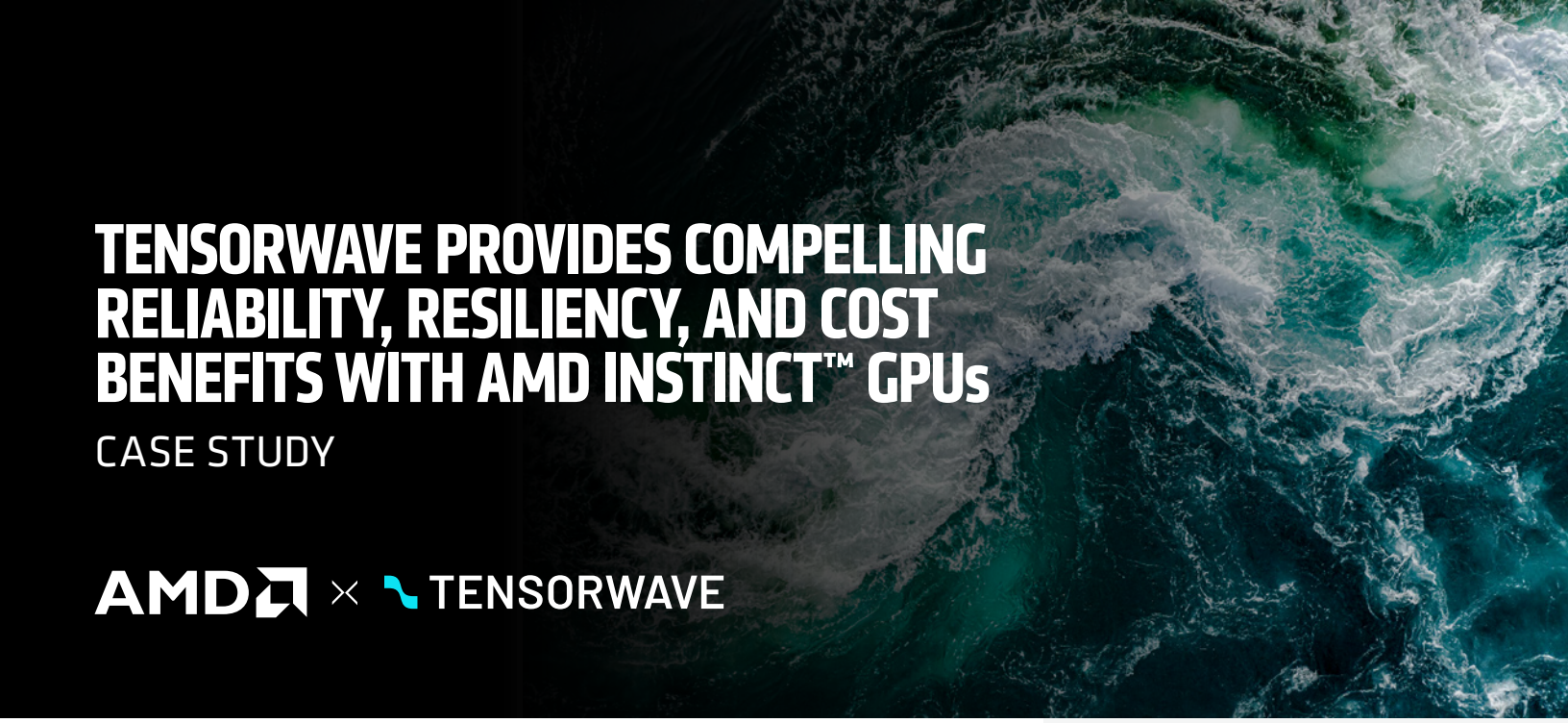


TENSORWAVE PROVIDES COMPELLING RELIABILITY, RESILIENCY, AND COST BENEFITS WITH AMD INSTINCT™ GPUs

CASE STUDY

AMD × TENSORWAVE

TensorWave is the largest AMD Instinct™ GPU-exclusive cloud, offering a next generation cloud designed to address the limitations of general purpose offerings. The company delivers specialized, reliable, and resilient infrastructure built for the demands of artificial intelligence (AI) and high performance computing (HPC), enabling customers to train, fine tune, and deploy complex AI models at scale.

CHALLENGE

The rise of AI is pushing the capabilities of generalized cloud service providers to the limit. Organizations betting their futures on AI innovation can no longer afford to rely on commodity infrastructure designed for yesterday's applications and databases. At the same time, AI innovation has been constrained by a single vendor ecosystem, where choice is restricted, supply is unpredictable, and costs escalate with each generation. These dynamics create AI roadblocks for organizations of all types, from startups seeking affordable access to specialized AI compute to global enterprises scaling AI initiatives.

Recognizing this need in the marketplace, TensorWave set out to build a specialized AI cloud free from the limitations of a single compute option without compromising on performance.

SOLUTION

To address the limitations of generalized clouds and the drawbacks of vendor lock in, TensorWave chose to build a specialized AI cloud, with GPU power provided exclusively by AMD Instinct™ GPUs and the open stack AMD ROCm™ software. This decision enables TensorWave to deliver reliable infrastructure designed for the realities of modern AI without the cost premiums and constraints of legacy ecosystems.

AMD Instinct™ GPUs offer the performance, scalability, and openness needed to redefine AI cloud economics. By focusing exclusively on AMD, TensorWave can design scalable and memory optimized clusters, so customers can train, fine tune, and serve models at unprecedented speed and efficiency. High density, liquid cooled systems with deterministic caching and massive bandwidth give customers the power to run mission critical AI workloads at scale.

Industry

Service provider

Challenges

Creating an open, reliable, and resilient AI cloud exclusively with AMD GPUs with an equal or better performance and lower cost than proprietary solutions

Technology Solution

- **TensorWave AI Cloud is powered exclusively by AMD Instinct™ GPUs**
- **AMD ROCm™ software**
- **Enhanced GPU portability provided by Modular MAX**

Results

Customers can port workloads to AMD Instinct™ GPUs in just minutes, and Modular has demonstrated up to 2X greater throughput when using AMD Instinct MI355X GPU clusters and Modular MAX, depending on model, as well as approximately 40-60% savings over NVIDIA B200 GPUs.[1]

A key factor in selecting AMD Instinct™ GPUs was creating an open option that allowed customers to avoid vendor lock-in. Moving workloads across hardware used to mean rewriting code, rebuilding pipelines, or sacrificing efficiency. TensorWave solved this by partnering with Modular, whose MAX platform makes portability simple, fast, and reliable. Drop-in containers let customers run the same inference commands across GPU vendors without code rewrites, complex migrations, or compromises. With Modular MAX and TensorWave, migrating to AMD Instinct™ GPUs takes just minutes—not months.

RESULTS

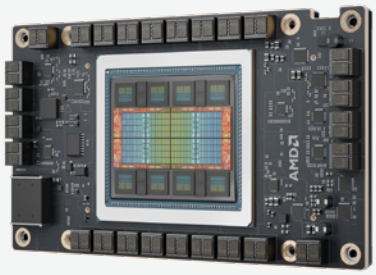
By choosing to implement on AMD Instinct™ GPUs exclusively, TensorWave created a cloud platform that delivers measurable, market defining advantages. The combination along with Modular MAX demonstrated up to 2x greater inference depending on the model and approximately 40–60% cost savings compared to NVIDIA B200 GPUs.[1] For TensorWave, these gains translate directly into differentiation.

Abstracting away the hardware gives TensorWave customers greater reliability, because infrastructure issues can be seamlessly resolved before they impact workloads. Faster inference means customers can deploy larger, more complex models without sacrificing latency. Smarter economics means organizations can scale AI workloads with confidence, knowing that performance aligns with predictable cost structures. And critically, Modular MAX helps customers move workloads in just minutes to tap into smarter scale and avoid the compromises of vendor lock in.

Together, these benefits allow TensorWave to position its AMD native cloud as a reliable, resilient, cost efficient, high performance, and easy to adopt alternative to general purpose cloud providers.

AMD Technology at a Glance

AMD Instinct™ MI355X, MI325X, and MI300X GPUs on large-scale training clusters expertly designed by TensorWave to maximize GPU performance.



LEARN MORE

Sign up to receive our data center content:
amd.com/instinct

ABOUT AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and X pages.

ABOUT TENSORWAVE

TensorWave is the AI and HPC cloud built for scale, speed, and serious workloads. Powered exclusively by AMD Instinct™ Series GPUs, TensorWave delivers high-bandwidth, memory-optimized infrastructure tailored for today's most demanding models—training or inference. Its supercomputing-class platform levels the playing field, giving every team access to the performance needed for their most ambitious AI initiatives.

[1] TensorWave, Real AI Workloads on AMD GPUs: Inference, Training, and Scaling, December 2025.