

# COST-EFFECTIVE GenAI PERFORMANCE

## CASE STUDY

AMD Instinct™ MI325X GPUs on Vultr help LiquidMetal AI deliver cloud-scale inference with efficiency and reach



LiquidMetal AI, founded in 2024 in San Francisco, developed Raindrop, an infrastructure layer designed to simplify high-performance GPU computing through standard APIs and the model context protocol (MCP).

Raindrop simplifies the management of cloud GPUs, allowing developers to deploy inference pipelines, coding assistants, and multi-agent systems without requiring direct hardware or kernel management.

Each deployment uses optimized inference templates—comparable to inference microservice modules—that enable applications to access GPUs through straightforward API calls. The platform addresses three common challenges in enterprise GPU adoption: observability, reproducibility, and multi-agent coordination.

With Raindrop, developers gain fine-grained visibility into model and agent behavior, tools for consistent performance across versions, and mechanisms such as SmartMemory and annotations that support context sharing across hardware environments.

By reducing operational overhead, Raindrop provides a structured approach to achieving performance, flexibility, and cost efficiency in modern AI workloads.

### THE INFRASTRUCTURE CHALLENGES BEHIND THE SCENES

Raindrop is designed to simplify AI-native development, allowing developers to launch new applications, such as knowledge assistants, generative agents, and complex backend services, without directly managing DevOps, provisioning, or scaling. For end users, the infrastructure remains invisible, but delivering a seamless experience requires precision and reliable performance.

Raindrop supports real-time, multi-tenant workloads with high throughput demands. Each customer interaction demands low-latency responsiveness, often exceeding 150 requests per second. With users distributed worldwide, LiquidMetal AI needs to maintain consistent performance across multiple regions.

**“Our platform can stand up authenticated, production-ready model APIs using templates built for AMD hardware. It’s the full-stack approach to deployment that developers and enterprises actually need, from prototype to production in minutes.”**

—Geno Valente,  
Head of Go-to-Market  
and Engineering



Throughput and cost efficiency are equally important. The platform has to process large volumes of concurrent requests while maximizing GPU utilization and keeping latency predictable. This is especially important for persistent agents and multi-turn interactions, where infrastructure costs accumulate quickly. LiquidMetal AI requires a solution that strikes a balance between performance, reliability, and cost efficiency.

Before adopting Vultr, Raindrop was deployed across several hyperscalers and content delivery networks (CDNs). As usage expanded, operational complexity and unpredictable costs became limiting factors. The team began seeking a unified approach that could deliver consistent performance at scale while preserving Raindrop's serverless design principles.

### **LIQUIDMETAL AI BUILDS ON VULTR WITH AMD INSTINCT GPUs**

To support real-time, high-throughput inference at global scale, LiquidMetal AI adopted Vultr cloud infrastructure powered by AMD Instinct™ MI325X GPUs. This combination provides the performance, flexibility, and control required for Raindrop's workloads while maintaining its serverless design principles.

Raindrop operates on AMD-based instances in an edge-integrated architecture designed for inference-heavy, latency-sensitive workloads. With eight GPUs per server, it efficiently scales agent workloads and supports multi-tenant deployments at high volumes. The architecture sustains throughput for models ranging from 70 billion to 700 billion parameters, including use cases such as SmartBucket retrieval—which provides natural language search over stored documents without separate indexing—and multi-agent orchestration. It also reduces batch inference costs for large-scale data processing, improving cost-per-token economics.

AMD Instinct MI325X GPUs are optimized for inference price-performance. Each GPU includes 256 GB of HBM3E memory and 6 TB/s of bandwidth, supporting longer context windows and larger batch sizes. These capabilities enable Raindrop to maintain low latency while maximizing GPU utilization and minimizing cost per token.

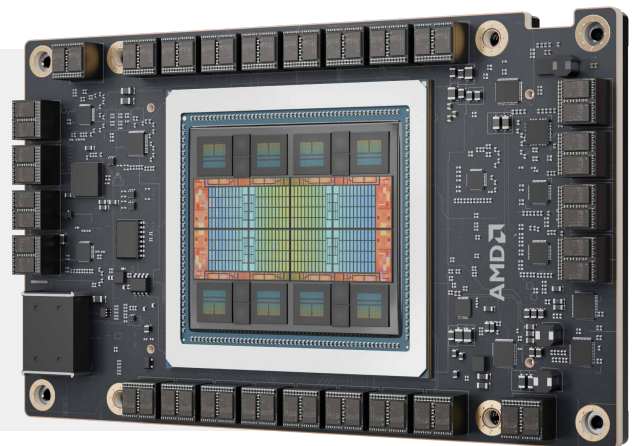
LiquidMetal AI leverages the AMD ROCm™ software stack for low-level performance tuning and optimization. The open ecosystem enables the faster integration of new agent architectures and memory workflows, allowing the team to iterate quickly and maintain control over how inference workloads are orchestrated without waiting on proprietary pipelines.

By using Vultr's global footprint and developer-oriented APIs, LiquidMetal AI can deploy AMD-powered infrastructure close to end users. This edge-integrated approach helps preserve responsiveness while expanding reach across regions.

Together, Vultr's distribution network, AMD Instinct GPUs, and Raindrop's API surface form a unified environment for inference workloads. The model demonstrates how open GPU ecosystems can support scalable, cost-efficient AI-native development across diverse applications.

**“The AMD Instinct MI325X delivered throughput that surprised us, performing 20 to 30 percent better on our workloads than competing options.”**

—Geno Valente, Head of Go-to-Market and Engineering



**THE LIQUIDMETAL AI ADVANTAGE:  
POWERED BY AMD AND VULTR**

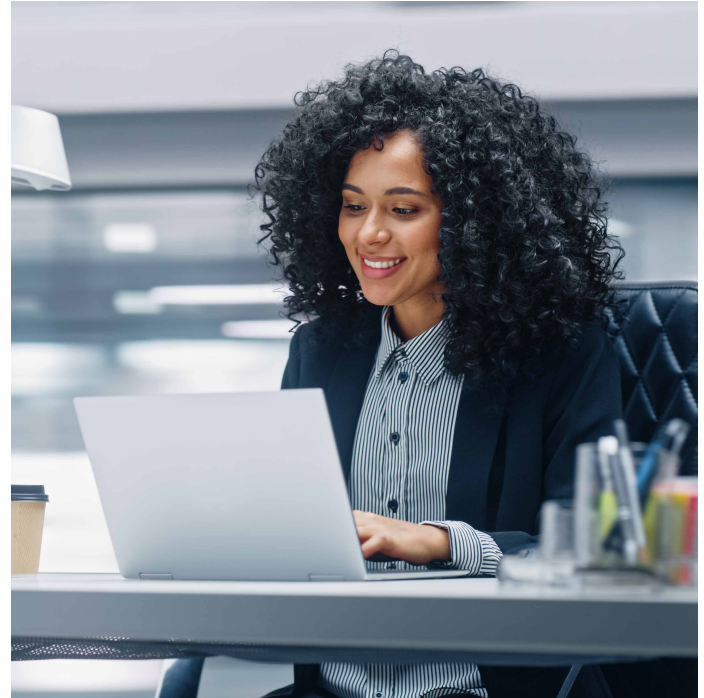
By building on Vultr’s global infrastructure and inference-optimized AMD Instinct™ GPUs, LiquidMetal AI can deliver scalable, cost-efficient cloud-native applications that meet the needs of modern GenAI. The infrastructure enables LiquidMetal AI to simplify development, accelerate deployment, and support the next generation of intelligent, multi-agent systems worldwide.

With Vultr and AMD Instinct GPUs, LiquidMetal AI achieves:

- **Lower cost per token** on inference workloads
- **Faster time to market** through immediate GPU availability
- **Reliable deployment** across 32 global cloud data center regions

This combination of affordability and reach allows LiquidMetal AI to scale Raindrop quickly. Developers see faster responses, steadier agent behavior, and lower cost. Every call to Raindrop benefits from AMD performance paired with LiquidMetal AI orchestration, bringing production-grade AI within reach for any team.

For end users—developers building assistants, agents, or research tools—this infrastructure delivers consistent performance and predictable economics. Whether they’re using Raindrop code, a customer-support bot, or a multi-agent research platform, they benefit from the performance and reach of AMD-powered inference running on Vultr’s global cloud network.



**“We chose AMD because it fits our workload: high memory, strong throughput, and open software.”**

–Geno Valente, Head of Go-to-Market and Engineering



**LEARN MORE**

Sign up to receive our data center content:  
[amd.com/instinct](https://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 VULTR**

Vultr is on a mission to make high-performance cloud infrastructure easy to use, affordable, and locally accessible for enterprises and AI innovators around the world. Vultr is trusted by hundreds of thousands of active customers across 185 countries for its flexible, scalable, global Cloud Compute, Cloud GPU, Bare Metal, and Cloud Storage solutions. In December 2024, Vultr announced an equity financing at a \$3.5 billion valuation. Founded by David Aninovsky and self-funded for over a decade, Vultr has grown to become the world’s largest privately held cloud infrastructure company.

**DISCLAIMERS**

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD’s Standard Terms and Conditions of Sale. GD-18u.

**COPYRIGHT NOTICE**

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Instinct, ROCm, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.