VULTR

DATASHEET

Vultr Cloud GPU, Powered by AMD Instinct™ MI325X and MI300X GPUs

Experience powerful and efficient AI and HPC deployments with exceptional memory capacity and bandwidth, and inference-optimized acceleration.

VULTR.COM

VULTR

Vultr Cloud GPU, Powered by AMD Instinct™ MI325X and MI300X GPUs

AMD Instinct™ MI325X and MI300X GPUs set new standards for powerful and efficient AI and HPC deployments. Coupled with the AMD ROCm™ open software ecosystem, the AMD Instinct™ MI325X and MI300X GPUs on the Vultr Cloud Platform ensure business can deploy a powerful and cost-effective platform without fearing vendor lock-in.

Why it's important right now

GPU acceleration has made AI's transformational effects possible, and AI is now used across every industry. In the initial stage, model training was the focus, with GPUs accelerating generative AI and deep learning model development. Now, as focus has begun to shift to put equal weight on model training and model deployment, a new paradigm is needed, one oriented towards efficient and cost-effective production deployments.

GPU clusters require significant energy and resources to run and scale, which results in considerable costs. Adding to the challenge, GPU clusters designed for model training are frequently not optimized to deliver production AI model deployments most efficiently.

The AMD Instinct™ MI325X and MI300X GPUs are powerful GPUs designed to excel at model training and be inference-optimized from the start. With exceptional memory capacity and bandwidth, AMD Instinct™ MI325X and MI300X GPUs can run larger models directly in memory, reducing the number of GPUs required and accelerating performance while reducing cost.

With the included AMD ROCm™ open software ecosystem, developers can leverage open standards and a proven software stack that simplifies programmability and portability. The ROCm™ open software ecosystem platform for GPU computing ensures freedom from vendor lock-in and limited architectural options.

Use cases

AI Inference

As AI training models have grown dramatically, so have the resulting AI Inference models. The AMD Instinct™ MI325X and MI300X GPUs have noteworthy memory capacity, ensuring larger models can be run on fewer GPUs for power efficiency.

With 192 GB of HBM3 memory and 5.3 TB/s memory bandwidth, the AMD Instinct™ MI300X GPU reduces required GPUs and accelerates performance. The AMD Instinct™ MI325X GPU, featuring 256 GB of HBM3E memory and 6.0 Tb/s memory bandwidth, further pushes the boundaries of what's possible in AI inference.

AI Model Training

The AMD Instinct™ MI325X and MI300X GPUs deliver the performance required for AI and deep learning model training. Their high memory bandwidth reduce model training time.

High-Performance Computing (HPC)

HPC is a fundamental requirement for many use cases requiring modeling, discovery, and prediction. The AMD Instinct™ MI325X and MI300X GPUs provide more raw acceleration performance in a standard platform, unlocking next-generation memory-intensive workloads.

Specifications

AMD Instinct™ MI325X GPU	
Form factor	OAM module
GPU compute units	304
Stream processors	1216
Peak engine clock	19,456
Memory capacity	Up to 256 GB HBM3E
Memory bandwidth	6.0 TB/s max. peak theoretical
Memory interface	8192 bits
AMD Infinity Cache™ (last level)	256 MB
Memory clock	Up to 6.0 GT/s

AMD Instinct™ MI300X GPU	
Form factor	OAM module
GPU compute units	304
Stream processors	1216
Peak engine clock	19,456
Memory capacity	Up to 192 GB HBM3
Memory bandwidth	5.3 TB/s max. peak theoretical
Memory interface	8192 bits
AMD Infinity Cache™ (last level)	256 MB
Memory clock	Up to 5.2 GT/s

AI peak theoretical performance		With sparsity
TF32 (TFLOPs)	653.7	1307.4
FP16 (TFLOPs)	1307.4	2614.9
BFLOAT16 (TFLOPs)	1307.4	2614.9
INT8 (TOPS)	2614.9	5229.8
FP8 (TFLOPs)	2614.9	5229.8

AI peak theoretical performance		With sparsity
TF32 (TFLOPs)	653.7	1307.4
FP16 (TFLOPs)	1307.4	2614.9
BFLOAT16 (TFLOPs)	1307.4	2614.9
INT8 (TOPS)	2614.9	5229.8
FP8 (TFLOPs)	2614.9	5229.8

HPC peak theoretical performance (TFLOPS)	
FP64 vector	81.7
FP32 vector	163.4
FP64 matrix	163.4
FP32 matrix	163.4

HPC peak theoretical performance (TFLOPS)	
FP64 vector	81.7
FP32 vector	163.4
FP64 matrix	163.4
FP32 matrix	163.4

Get started today

Discover the benefits of using Vultr Cloud GPU, powered by AMD Instinct™ MI325X and MI300X GPUs, for your next project. Visit vultr.com to learn more about our powerful lineup of AMD GPUs.

Learn more about
Vultr Cloud GPU

Contact us at vultr.com to get started. →