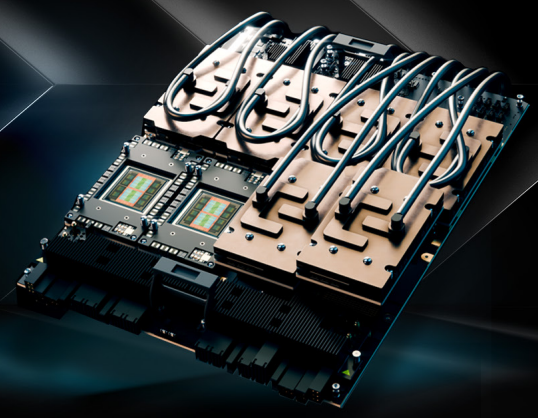
AMD INSTINCT™ MI355X PLATFORM

ADVANCED DIRECT LIQUID COOLED GPU SOLUTION FOR AI, INFERENCE, AND TRAINING



AMD INSTINCT

ULTIMATE AI & HPC PERFORMANCE

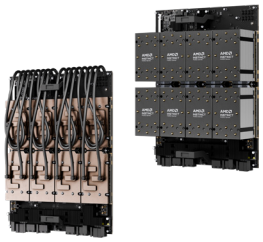
Built on the cutting-edge 4th Gen AMD CDNA™ architecture, AMD Instinct™ MI355X Platform features eight GPUs on a Universal Base Board (UBB) form factor. It features powerful and energy-efficient cores with excellent performance per watt to drive the next era of AI and HPC innovation.

With exceptional efficiency and performance, the platform excels at high-speed inference, training massive AI models, and tackling the most complex computational challenges—from scientific simulations to large-scale data processing and advanced modeling.

DIRECT LIQUID COOLED DESIGN

The global adoption of AI technology demands ever greater amounts of processing capacity that can only be achieved by reaching for higher densities and more efficient approaches to power and cooling. While an air-cooled option is available, the direct liquid cooling option enables AMD Instinct MI355X GPUs to consume up to 1400W. Liquid cooling helps reduce GPU die temperatures, expanding the accelerator's power envelope, facilitating higher clock speeds and delivered performance. Other component temperatures, for example the PCIe Gen 5 interfaces to the host system, are moderated by air-cooled heat sinks.

Supplied with bare GPU dies, server vendors install cold plates from one of two qualified vendors. This enables servers to use cooling technology they engineer. The use of cold plates instead of heat sinks for the individual GPUs enables the UBB to be integrated into standard and open compute servers, helping customers achieve higher densities in direct liquid cooled and air-cooled configurations.



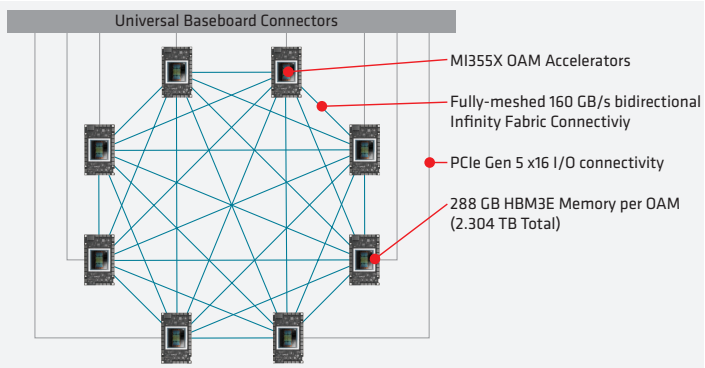
AI PEAK THEORETICAL PERFORMANCE		W/SPARSITY
FP16 (PFLOPS)	20.1	40.3
BFLOAT16 (PFLOPS)	20.1	40.3
INT8 (PFLOPS)	40.3	80.5
INT4 (POPS)	40.3	80.5
FP8 (PFLOPS)	40.3	80.5
FP6 (PFLOPS)	80.5	161.1
FP4(PFLOPS)	80.5	161.1

HPC PEAK THEORETICAL PERFORMANCE	
FP64 VECTOR (TFLOPS)	628.8
FP32 VECTOR (PFLOPS)	1.3
FP64 MATRIX (TFLOPS)	628.8
FP32 MATRIX (PFLOPS)	1.3

VIDEO DECODERS AND VIRTUALIZATION	
DECODERS†	32 groups for HEVC/H.265, AVC/H.264, V1, or AV1
JPEG/MJPEG CODEC	320 cores, 10 cores per group
GPU PHYSICAL PARTITIONS	SR-IOV, up to 64 partitions
MEMORY PARTITIONS	1 or 4 per module

SPECIFICATIONS	
FORM FACTOR	UBB 2.0 module
COOLANT	PG-25
MAXIMUM LIQUID INLET TEMPERATURE	43°C
RECOMMENDED FLOW RATE	2.1 l/min per OAM
LITHOGRAPHY	TSMC 3nm/6nm FinFET
GPU COMPUTE UNITS	2048
MATRIX CORES	8192
STREAM PROCESSORS	131,072
PEAK ENGINE CLOCK	2200 MHz
MEMORY CAPACITY	2.3 TB HBM3E
MEMORY BANDWIDTH	8 TB/s per GPU max. peak theoretical
MEMORY INTERFACE	8192 bits
AMD INFINITY CACHE™ (LAST LEVEL)	256 MB (per GPU)
SCALE-UP AMD INFINITY FABRIC™ LINKS	7 (per GPU) x 144 GB/s
I/O TO HOST CPU	8 PCIe® Gen 5 x16 (128 GB/s)
RAS FEATURES	Full-chip ECC memory, page retirement, page avoidance
MAXIMUM TBP	1400W per module

†Video codec acceleration (including at least the HEVC (H.265), H.264, VP9, and AV1 codecs) is subject to change and not operable without inclusion/installation of compatible media players. GD-176



AMD Instinct MI355X Platform Architecture

INDUSTRY-STANDARD FORM FACTOR

The platform combines the power of eight accelerators on an industry-standard universal baseboard (UBB 2.0). The eight Open Compute Project (OCP) Accelerator Modules (OAMs) are connected with an AMD Infinity Fabric™ mesh that provides direct connectivity between each of the GPUs over 160 GB/s bidirectional links. Each MI355X connects with its peers through seven links, plus one PCIe® Gen 5 x16 connection per OAM device for upstream server and/or I/O connectivity. Each accelerator boasts 288 GB of HBM3E memory with 8 TB/s of memory bandwidth. This combines into a massive 2.3 TB of coherent, shared memory.

BREAKTHROUGH AI ACCELERATION

With the new expanded FP6 and FP4 datatype support, AMD Instinct MI355X GPUs maximize computational throughput, memory bandwidth utilization, and energy efficiency, enabling faster, more power-efficient AI inference compared to previous-generation datatype support. Enhanced FP16 and FP8 processing, combined with added next-gen FP6 and FP4 capabilities, position the AMD Instinct MI350 Series products to deliver exceptional performance for advanced generative AI models—pushing the boundaries of AI acceleration.

SEAMLESS GPU UPGRADES FOR BETTER PERFORMANCE1

The AMD GPU Operator simplifies deployment and management of AMD Instinct GPUs in Kubernetes clusters, helping enable effortless configuration of GPU-accelerated workloads, streamlining operations while accelerating time to market.

Scaling software from prior-generation, air-cooled accelerators is simplified by the AMD ROCm™ platform. Day-0 support enables optimized models to be available upon product release.

ADVANCED SECURITY FOR AI & HPC DEPLOYMENTS

AMD Instinct MI355X GPUs deliver robust security to protect AI models, data, and system integrity. They help ensure only trusted firmware runs, verify hardware authenticity, enable secure multi-tenant GPU sharing, and encrypt high-speed GPU communication. These protections help enhance reliability, scalability, and data confidentiality, making them ideal for cloud AI, enterprise, and mission-critical workloads.

SCALABLE, FUTURE-PROOF NETWORKING

Ethernet-based AI and HPC networking enables massive hyper-class scalability, low TCO, and eliminates vendor lock-in, helping ensure open, flexible AI clusters.

OPTIMIZED SOFTWARE & FRAMEWORK INTEGRATION

The foundation of AMD accelerated computing, AMD ROCm software empowers AI developers to fully leverage AMD Instinct GPUs including the latest Instinct MI350 Series for both inference and training. With Day-0 support for major AI frameworks like PyTorch, TensorFlow, JAX, and ONNX Runtime, the latest advancements in ROCm helps simplify AI model migration and deployment, optimizing hardware efficiency while minimizing code changes. Through strategic collaborations with AI leaders such as OpenAI, PyTorch, Hugging Face, Databricks and more, ROCm delivers high-performance, out-of-the-box AI solutions, enabling smooth integration, and robust partner support.

The ROCm 7.0 release delivers significant advancements in AI training and infererence. Mixed-precision training is optimized through support for FP16, FP8, BF16, FP6, FP4 datatypes, helping accelerate training while maintaining high accuracy for AI models. Fast data loading and augmentation pipelines can reduce CPU-GPU bottlenecks, helping ensure high-throughput data preprocessing for large datasets. Once models are trained, AI inference can be optimized with low-latency kernels to help enhance transformer-based model execution, accelerating attention mechanisms for fast response times. With enhanced quantization support, ROCm software is designed to improve FP8, FP6, and FP4 inference, enabling efficient low-bit AI processing that reduces power consumption while maintaining accuracy., helping enable real-time processing, ultra-low latency, and increased power efficiency.

TRUSTED BY AI LEADERS

AI giants including Meta have chosen AMD Instinct™ GPUs for their own large-scale AI deployments, powering models like Llama 405B and GPT.

With adoption across many leading and emerging cloud service providers, original equipment manufacturers, and original design manufacturers, AMD is driving next-gen AI at scale.

LEARN MORE

The AMD Instinct MI355X Platform is available through AMD solution partners. Standard form factors facilitate adoption into enterprise servers so that you can use the same power in your data center, or in the cloud from the leading superscalars. Learn more at AMD.com/INSTINCT.

1. MI355X Platform - PID#253461436

Footnote explanations are available at: https://www.amd.com/en/legal/claims/instinct.html

LE-92704-00 5/25