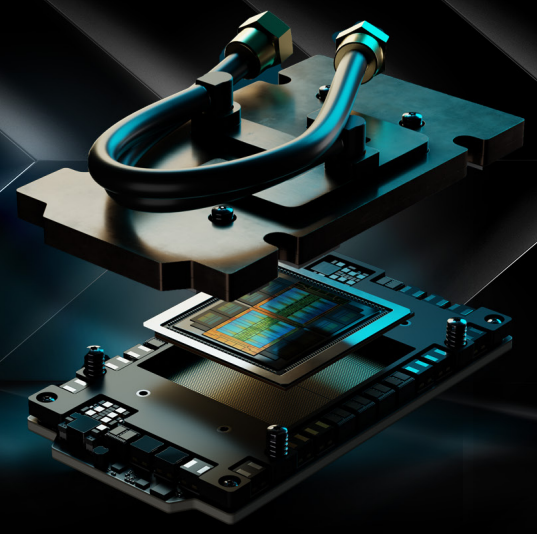
AMD INSTINCT™ MI355X GPU

LEADING-EDGE GPU FOR GENERATIVE AI, INFERENCE, TRAINING, AND HIGH PERFORMANCE COMPUTING



AMD INSTINCT

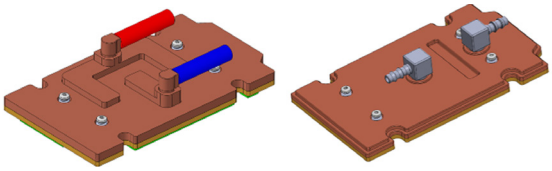
LEADERSHIP AI AND HPC ACCELERATION

The AMD Instinct™ MI350 Series (featuring both MI350X and MI355X GPUs) sets a new standard for generative AI and high performance computing (HPC) in data centers. Built on the new cutting-edge 4th Gen AMD CDNA™ architecture, the accelerator delivers exceptional efficiency and performance for training massive AI models, high-speed inference, and complex HPC workloads including scientific simulations, data processing, and computational modeling.

DESIGNED FOR HIGH-DENSITY COMPUTING

The AMD Instinct MI355X GPU is purpose built for high density computing environments, offering more peak performance than its counterpart the MI350X. This increased power enables the MI355X to sustain higher performance over time, minimizing throttling and maximizing throughput during prolonged or intensive workloads.

When deployed in an 8-GPU AMD Instinct MI355X platform, the solution uses the AMD Universal Base Board (UBB 2.0), which can be integrated into server designs as compact as 2U. The platform connects to the host through standard connectors, making it ideal for space constrained, performance driven deployments.



SEAMLESS SCALABILITY & DEPLOYMENT

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.

AI PEAK THEORETICAL PERFORMANCE		W/SPARSITY
FP16 VECTOR (TFLOPS)	157.3000	N/A
FP16 MATRIX (PFLOPS)	2.5166	5.0332
BFLOAT16 (PFLOPS)	2.5166	5.0332
INT8 MATRIX (POPS)	5.0332	10.0664
MXFP8 (PFLOPS)	5.0332	N/A
OCP-FP8 (PFLOPS)	5.0332	10.0664
MXFP6 (PFLOPS)	10.0663	N/A
MXFP4 (PFLOPS)	10.0663	N/A

HPC PEAK THEORETICAL PERFORMANCE	
FP64 VECTOR (TFLOPS)	78.6
FP32 VECTOR (TFLOPS)	157.3
FP64 MATRIX (TFLOPS)	78.6
FP32 MATRIX (TFLOPS)	157.3

DECODERS AND VIRTUALIZATION	
DECODERS†	4 groups for HEVC/H.265,AVC/H.264, VP9, or AV1
JPEG/MJPEG CODEC	40 cores, 10 cores per group
GPU PHYSICAL PARTITIONS	Up to 8 @ 36 GB
MEMORY PARTITIONS	1 or 4

SPECIFICATIONS	
FORM FACTOR	OAM module
LITHOGRAPHY	TSMC 3nm/6nm FinFET
I/O DIES (IODS)	2 mirrored
GPU COMPUTE UNITS	256
MATRIX CORES	1024
STREAM PROCESSORS	16,384
PEAK ENGINE CLOCK	2.4 GHz
MEMORY CAPACITY	288 GB HBM3E
MEMORY BANDWIDTH	8 TB/s
MEMORY INTERFACE	8192 bits
AMD INFINITY CACHE™ (LAST LEVEL)	256 MB
SCALE-UP AMD INFINITY FABRIC™ LINKS	7x 153.6 GB/s
I/O INTERCONNECT	1 PCIe® Gen 5 x16 (128 GB/s)
RAS FEATURES	Full-chip ECC memory, page retirement, page avoidance
MAXIMUM TBP	1400W

†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

NEXT-GEN COMPUTE POWER WITH EXPANDED DATATYPE SUPPORT

With expanded MXFP6 and MXFP4 datatype support, the AMD Instinct MI355X GPU maximizes computational throughput, memory bandwidth utilization, and energy efficiency, enabling faster, more power-efficient AI inference compared to previous-generation accelerators. Enhanced FP16, MXFP8, and OCP-FP8 processing, combined with added next-gen MXFP6 and MXFP4 capabilities, position the AMD Instinct MI355X to deliver exceptional performance for advanced AI models—pushing the boundaries of AI acceleration.

Featuring a massive 288 GB HBM3E memory capacity and 8 TB/s bandwidth, the GPU provides exceptional AI capabilities handling larger models with fewer accelerators. These innovations help reduce server resource requirements, promote easy scaling and management of AI workloads, and can help lower total cost of ownership (TCO) for AI-driven data centers.

BUILT-IN SECURITY FOR AI & HPC DEPLOYMENTS

Security is essential for AI and HPC. The AMD Instinct™ MI350 Series integrates advanced security to protect AI models, data, and system integrity. Device Secure Boot and Secure Update & Recovery help ensure only trusted firmware runs, while Platform-Level DICE Identity & Attestation verifies GPU authenticity to prevent unauthorized access.

For multitenant AI and HPC environments, SR-IOV helps enable secure, efficient GPU resource sharing across multiple virtual machines while maintaining isolation between tenants. AMD Infinity Fabric™ Link security helps protect high-speed GPU-to-GPU communication. These features help enhance reliability and trust, making AMD Instinct GPUs an excellent choice for cloud AI, enterprise, and mission-critical workloads in finance, healthcare, and government.

The AMD Instinct MI355X supports massive Ethernet-based AI networking with hyper-class scalability, low costs, and no vendor lock-in for open, flexible AI infrastructure.

OPEN AND OPTIMIZED AI SOFTWARE STACK

Built on the AMD commitment to open-source innovation, AMD Instinct MI350 Series GPUs integrate with the next-generation AMD ROCm™ software stack—the industry's premier open alternative for AI and HPC. The ROCm platform supports all major AI and HPC frameworks, inference engines, and model-serving systems including PyTorch, TensorFlow, JAX, ONNX Runtime, Kokkos, Raja, SGLang, Triton, and vLLM, enabling effortless model deployment with minimal code changes and maximum flexibility.

The latest ROCm software enhancements further optimize AI inference, training, and framework compatibility, delivering high throughput and ultra-low latency for demanding workloads such as natural language processing (NLP), computer vision, and beyond. Through strategic collaborations with AI leaders such as OpenAI, Meta, PyTorch, Hugging Face, Databricks, and Lamini, the smooth integration the ROCm platform provides enables developers and businesses to accelerate AI inference and training with confidence, unlocking faster innovation and deployment.

TRUSTED BY AI LEADERS

Microsoft and Meta, among other leading innovators, trust AMD Instinct™ GPUs for large-scale AI, powering models like Llama 405B and GPT. Broad AMD Instinct GPU adoption by CSPs & OEMs are helping to drive next-gen AI at scale.

Collaborations between AMD and leading cloud service providers (CSPs), original equipment manufacturers (OEMs), and platform designers drive a robust ecosystem of AMD Instinct MI350 Series-powered servers, delivering a comprehensive and diverse portfolio of AI and HPC solutions to the market.

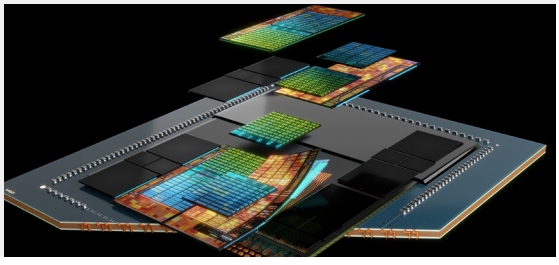
LEARN MORE

For more information, visit AMD.com/INSTINCT.

MULTI-CHIP ARCHITECTURE

The MI355X uses the 4th Gen AMD CDNA™ multi-chip architecture based on 3nm process technology to enable dense compute and high-bandwidth memory integration. Each OAM module includes:

- Eight accelerated compute dies (XCDs) with 32 compute units (CUs), 32 KB of L1 cache per CU, 4 MB shared L2 cache shared across CUs, and 256 MB of AMD Infinity Cache™ shared across 8 XCDs. The compute units support a broad range of precisions for both AI/ML and HPC acceleration, native hardware support for sparsity, and enhanced computational throughput.
- Four supported decoders for HEVC/H.265, AVC/H.264, VP9, or AV1, each with an additional 40-core JPEG/MPEG CODEC
- 288 GB of HBM3E memory with 8 TB/s on-package peak throughput
- SR-IOV for up to 8 partitions



COHERENT SHARED MEMORY

AMD Instinct accelerators facilitate large models with hybrid hardware/software memory coherency between all eight accelerators on a universal baseboard with 160 GB/s bidirectional bandwidth between each GPU to accelerate memory-intensive AI, ML, and HPC models.

1. MI355X GPU - PID#253461434A

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

LE-92703-00 10/25