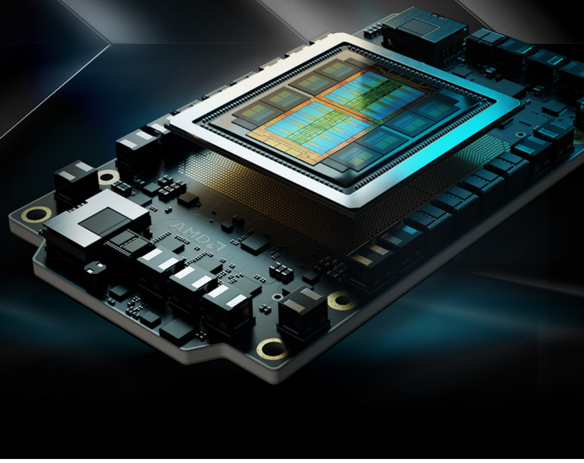
AMD INSTINCT™ MI350X GPU

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



LEADERSHIP AI AND HPC ACCELERATION

The AMD Instinct MI350 Series GPUs (featuring both MI350X and MI355X GPUs) set 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, these GPUs deliver exceptional efficiency and performance for training massive AI models, high-speed inference, and complex HPC workloads including scientific simulations, data processing, and computational modeling. The MI350X is interconnected using AMD Infinity Fabric™ Link enabling high-bandwidth GPU-to-GPU communication while maintaining Universal Base Board (UBB 2.0) compatibility.

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.

To meet evolving customer and infrastructure demands, the new AMD Instinct™ MI350X GPU integrates seamlessly with prior-generation AMD Instinct MI300 Series platforms (including the Instinct MI300X and MI325X) and competitive infrastructures—offering cost-effective performance upgrades. For higher-density computing, the AMD instinct MI350 Series offers a full range of optimized cooling solutions, including air-cooled and direct liquid-cooled options, designed to support both compact deployments and high-capacity cooling configurations.

NEXT-GEN COMPUTE POWER WITH EXPANDED DATATYPE SUPPORT

With expanded FP6 and FP4 datatype support, Instinct MI350 Series GPUs maximize computational throughput, memory bandwidth utilization, and energy efficiency, enabling faster, more power-efficient AI inference compared to previous-generation accelerators. Enhanced FP16 and FP8 processing, combined with added next-gen FP6 and FP4 capabilities, position the AMD Instinct MI350 Series 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 AMD Instinct MI350 Series provides exceptional

AI PEAK THEORETICAL PERFORMANCE		W/SPARSITY
FP16 (PFLOPS)	2.3069	4.6138
BFLOAT16 (PFLOPS)	2.3096	4.6192
INT8 (PFLOPS)	4.6137	9.2274
INT4 (POPS)	4.6137	9.2274
FP8 (PFLOPS)	4.614	9.2274
FP6 (PFLOPS)	9.2275	18.455
FP4(PFLOPS)	9.2275	18.455

HPC PEAK THEORETICAL PERFORMANCE	
FP64 VECTOR (TFLOPS)	72.1
FP32 VECTOR (TFLOPS)	144.2
FP64 MATRIX (TFLOPS)	72.1
FP32 MATRIX (TFLOPS)	144.2

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
ACTIVE INTERPOSER DIES (AIDS)	2 mirrored
GPU COMPUTE UNITS	256
MATRIX CORES	1024
STREAM PROCESSORS	16,384
PEAK ENGINE CLOCK	2.2 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 144 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	1000W

†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

AI capabilities handling larger models with fewer GPUs. 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.

OPEN AND OPTIMIZED AI SOFTWARE STACK

Built on the AMD commitment to open-source innovation, AMD Instinct MI350 Series GPUs are seamlessly integrated 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, vLLM, and more—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.

AMD INSTINCT

Through strategic collaborations with AI leaders such as OpenAI, Meta, PyTorch, Hugging Face, Databricks, and Lamini, AMD ROCm software delivers Day-0 support, helping ensure AMD Instinct GPUs are optimized to run the latest AI models and frameworks immediately upon release. This smooth integration enables developers and businesses to accelerate AI inference and training with confidence, unlocking faster innovation and deployment.

SCALABLE, FUTURE-PROOF AI NETWORKING

AMD Instinct MI350 Series GPUs are built for next-gen Ethernet-based AI networking, enabling massive hyper-class scalability, low costs, and an open, flexible architecture—eliminating vendor lock-in and helping ensure seamless interoperability across multiple networking vendors.

As a founding member of the Ultra Ethernet Consortium (UEC), AMD helps shape the future of AI networking standards, making its solutions future-ready for large-scale AI clusters.

A STRONG ECOSYSTEM WITH INDUSTRY LEADERS

Marquee AI leaders like Meta trust AMD Instinct GPUs to power large-scale AI deployments, delivering leadership price-performance for inference and serving as the AI infrastructure choice for models like Llama 405B and GPT. Their adoption underscores proven performance, efficiency, and scalability, reinforcing the AMD position as a trusted supplier for next-generation AI.

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.

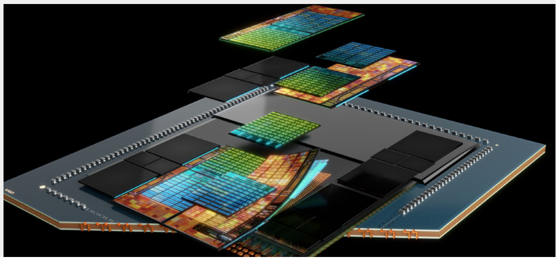
MULTI-CHIP ARCHITECTURE

The MI350X 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] MI350X GPU - PID#253461430

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