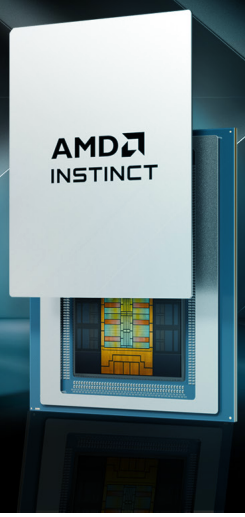
AMD INSTINCT™ MI455X GPU

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



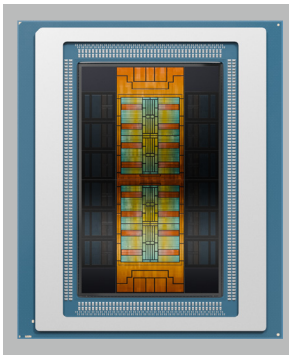
AMD INSTINCT

REDEFINE AI & HPC AT EVERY SCALE

The AMD Instinct™ MI455X GPU is engineered for the demands of modern artificial intelligence and high-performance computing workloads. Designed to scale across hyperscale data centers, sovereign AI deployments, and research environments, it combines the performance of 5th Gen AMD CDNA™ architecture with an open software foundation and standards-based system architecture. Advanced partitioning and virtualization capabilities allow models, tools, and operational practices to move seamlessly between deployment environments.

PERFORMANCE LEADERSHIP FOR AI WORKLOADS

The GPU delivers industry-leading low-precision data format performance—up to 4x the performance of the previous-generation AMD Instinct MI355X on 4-bit and 8-bit matrix floating-point operations.[MI400-006] Four-bit precision delivers peak performance of up to 40 PFLOPs; at eight-bit precision, peak performance exceeds 20 PFLOPs. This implementation adheres to the OCP MX standard, ensuring hardware-software interoperability across modern AI GPUs. Memory capacity reaches 432 GB of HBM4, with industry-leading peak theoretical bandwidth of 23.3 TB/s—up to 2.9x higher peak memory bandwidth than the previous-generation AMD Instinct MI355X GPU.[MI400-008]



The AMD Instinct MI455X GPU is based on 5th Gen AMD CDNA architecture in a multi-chip package

RACK-SCALE INFRASTRUCTURE FOR FRONTIER AI

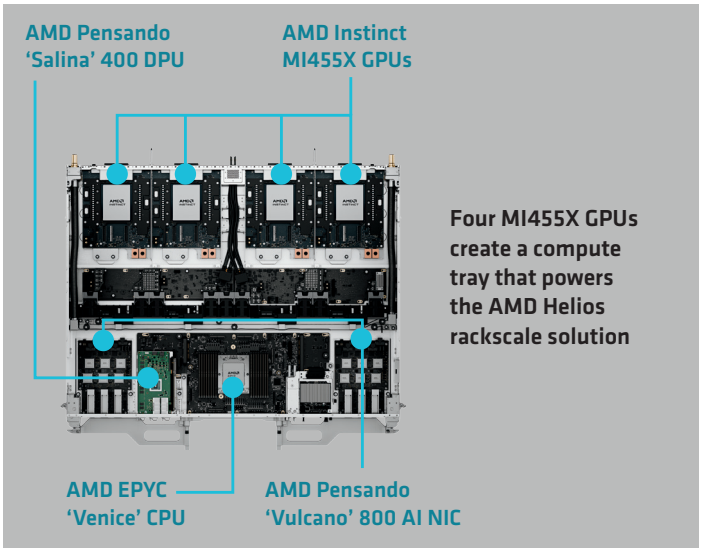
The next phase of AI requires complete rack-scale solutions designed for hyperscale deployment. The GPU is designed to be deployed in the AMD Helios rackscale solution with 72 GPUs per

AI PEAK THEORETICAL PERFORMANCE	TFLOPS	W/STRUCTURED SPARSITY
OCP MXFP4	40265	
OCP MXFP6	20133	
OCP MXFP8	20133	
OCP FP8	20133	
MATRIX FP16	5033	10066
VECTOR FP16	315	
MATRIX FP32	315	
VECTOR FP32	315	
MATRIX FP64	5	
VECTOR FP64	5	
MATRIX INT8	5033	10066
MATRIX BF16	5033	10066

DECODERS AND VIRTUALIZATION	
DECODERS†	4 engines for HEVC/ H.265, AVC/H.264, VP9, or AV1
JPEG/MJPEG CODEC	40 cores, 10 cores per group
VIRTUALIZATION SUPPORT	SR-IOV
MEMORY PARTITIONS	4

SPECIFICATIONS	
FORM FACTOR	EAM module
COOLING	Liquid
LITHOGRAPHY	TSMC 2nm/3nm
COMPUTE UNITS	256
ACCELERATED COMPUTE DIES (XCD)	8
I/O DIES (IOD)	2
PEAK ENGINE CLOCK	2.4 GHz
MEMORY CAPACITY	432 GB HBM4
PEAK MEMORY BANDWIDTH	23.3 TB/s
CPU TO GPU INTERCONNECT	256 GB/s
PEAK SCALE-UP UALoE BANDWIDTH	3.6 TB/s
PEAK SCALE-OUT BANDWIDTH	600 GB/s
RAS FEATURES	Full-chip ECC memory, page retirement

†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



Four MI455X GPUs create a compute tray that powers the AMD Helios rackscale solution

rack. Four liquid-cooled AMD Instinct MI455X GPUs are packaged on an enhanced accelerator module (EAM) form factor. One of these is installed in an AMD Instinct MI455X compute tray. Each GPU is accessed through AMD Infinity Fabric™ connections to an AMD EPYC™ "Venice" server CPU with 96 cores. In the AMD Helios solution, 18 compute trays are housed in a rack. The compute trays are interconnected by a switched scale-up fabric of 3.6 TB/s per GPU using UALoE (UAL over Ethernet.) Multiple links per GPU deliver consistent bandwidth and rack-wide GPU memory access in a single load/store domain. The fabric offers resilience with automatic failover in the event of errors or disabled links.

To support multi-rack datacenters, standards-based Ethernet–aligned with the Ultra Ethernet Consortium (UEC)–enables high-performance scaling across racks and clusters without proprietary fabrics. The compute tray powers the Ethernet fabric with AMD Pensando™ Vulcano AI NICs, delivering lossless, high-throughput networking with programmable pipelines and AMD ROCm™ software or platform Communication Collectives Library (RCCL) proxy acceleration. Open-source tools for telemetry, profiling, and cluster management round out a complete, standards-based scalable solution.

FLEXIBLE PARTITIONING AND MULTITENANT SUPPORT

The AMD Instinct MI455X GPU is designed to serve both dedicated and shared workloads with a hierarchy of partitioning capabilities. A 72-GPU AMD Helios rack can be subdivided into virtual pods that range in size from four GPUs sharing a common CPU within a single compute tray, up to the full complement of compute nodes in the rack. Compute nodes within a virtual pod do not need to be physically adjacent, enabling flexible allocation across the rack.

At the GPU level, the MI455X supports compute partitioning–the ability to subdivide a single GPU's compute units, memory, and video decoders into smaller, independent partitions. Complementing this, spatial memory partitioning divides GPU memory into isolated NUMA regions, enabling each partition to operate independently and optimizing memory access patterns for multitenant and parallel workloads. Together, these capabilities support high-volume inference, distributed training, and fine-tuning across multi-trillion-parameter AI models in shared and virtualized environments.

ADVANCED SECURITY AT SCALE

AMD Instinct MI455X GPUs provide advanced security including expanded device-level attestation, support for universal link encryption, memory encryption, and secure multi-GPU scaling designed for data and model confidentiality at scale. AMD Helios brings next-generation security enhancements for large AI training clusters and distributed inference deployments, enabling protected model execution across massively scaled GPU fabrics.

INDUSTRY'S PREMIER OPEN-SOURCE AI STACK

AMD ROCm software is the industry's premier open-source AI stack–validated in exascale-class AI and HPC environments and designed for a developer-first approach to enablement. Native support across the largest open-source projects and Day-0 model support ensure that the latest models and frameworks run on AMD hardware from day one.

Supported frameworks and runtimes include PyTorch, TensorFlow, JAX, ONNX Runtime, vLLM, SGLang, and DeepSeek. AMD Primus drives high-performance training and fine-tuning with distributed execution optimized for rack-scale and cluster-scale deployments. Combined with open rack-scale architecture and Ethernet-based networking, ROCm software enables seamless scale-up within racks and scale-out across clusters–giving developers and enterprises a unified software foundation for the full spectrum of AI and HPC workloads.

AMD is empowering an expanded set of developers through AMD AI Academy and the AMD Developer Cloud, accelerating the pace of innovation and broadening ecosystem partnerships across the AI community.

POWER THE NEXT WAVE OF FRONTIER AI

AMD Instinct GPUs are trusted by eight of the world's top ten AI organizations – including OpenAI and Meta–with public commitments to deploy up to six gigawatts of AMD Instinct GPU capacity as part of strategic, multi-year collaborations. With Oracle Cloud Infrastructure (OCI) planning to deploy 50,000 MI455X GPUs beginning in 2026, AMD Instinct GPUs continue to provide an open, scalable foundation for the world's most demanding AI workloads. These endorsements bestow confidence that the Instinct AMD MI455X GPU is ready for your largest deployments.

LEARN MORE

For more information, visit AMD.com/INSTINCT.

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

LE-93204-00 07/26