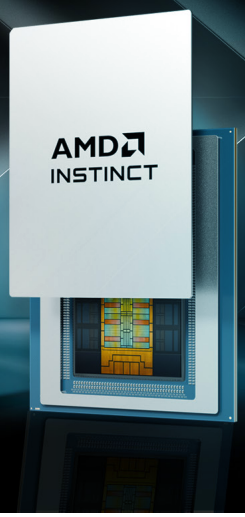
AMD INSTINCT™ MI450X GPU

FULL COMPUTE. FULL BANDWIDTH. MORE CHOICE.

BUILT FOR AI INFERENCE, TRAINING, AND FINE-TUNING



AMD INSTINCT

MORE CHOICE FOR AI AT SCALE

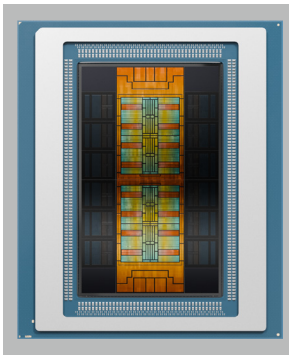
The AMD Instinct™ MI450X GPU expands the AMD Instinct MI400 Series portfolio with more choice for AI deployments on AMD Helios™ rackscale solution. It pairs 288 GB of HBM4 with the same core compute, memory bandwidth, and connectivity capabilities as the AMD Instinct MI455X GPU. Built on 5th Gen AMD CDNA™ architecture, it helps customers match memory capacity to their models, throughput targets, and response-time needs. A common rack-scale architecture and AMD ROCm™ software foundation support flexible deployment across AI workloads.

PERFORMANCE ALIGNED TO INFERENCE NEEDS

MI450X delivers full compute and memory bandwidth for workloads aligned with its 288 GB HBM4 capacity. Peak theoretical performance exceeds 40 PFLOPs at four-bit precision and 20 PFLOPs at eight-bit precision, with support for OCP MX formats and 23.3 TB/s of memory bandwidth. For interactive inference, customers can balance total throughput with the response speed their applications require, while exploring more favorable deployment economics.

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



The AMD Instinct MI450X GPU is based on 5th Gen AMD CDNA architecture in a multi-chip package

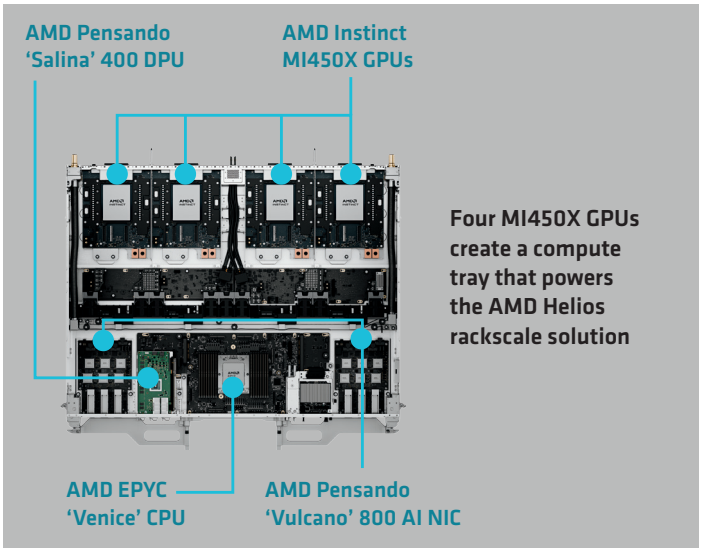
The MI455X GPU is ideal for applications where additional memory capacity is not a limiting factor. The right choice depends on the model, context length, concurrency, throughput, response-time targets, and overall deployment cost.

RACK-SCALE AI ON ONE AMD HELIOS PLATFORM

AMD Helios brings MI450X GPUs, AMD EPYC™ CPUs, and AMD Pensando™ networking together in a common rack-scale architecture designed for AI, with 72 GPUs per rack.

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	288 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 MI450X GPUs create a compute tray that powers the AMD Helios rackscale solution

Each compute tray combines four liquid-cooled AMD Instinct MI450X GPUs in the enhanced accelerator module (EAM) form factor. Each GPU connects through AMD Infinity Fabric™ technology to an AMD EPYC™ "Venice" server CPU with 96 cores. AMD Helios houses 18 compute trays in a rack. The trays are interconnected by a switched scale-up fabric of 3.6 TB/s per using Ultra Accelerator Link over Ethernet (UALoE). 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, supporting AI deployments at rack scale.

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™ 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 MI450X 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 MI450X 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 MI450X 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.

AN OPEN ECOSYSTEM FOR CUSTOMER NEEDS

AMD ROCm software gives MI450X and MI455X deployments a common open-source AI foundation. Flexible compute architectures and open networking standards enable customer-focused solutions through ecosystem partners. Customers can choose memory capacity to fit their needs while retaining familiar AI models, frameworks, and tools.

Supported frameworks and runtimes include PyTorch, TensorFlow, JAX, ONNX Runtime, vLLM, and SGLang. AMD Primus drives training and fine-tuning with distributed execution optimized for rack-scale and cluster-scale deployments. Open rack-scale architecture and Ethernet-based networking enable ROCm software to support scale-up within racks and scale-out across clusters. A shared hardware and software foundation helps reduce integration work, enabling AMD and its partners to bring new configurations to customers faster.

AMD AI Academy and the AMD Developer Cloud support developer enablement and ecosystem collaboration. Together, these resources help customers build expertise and bring AI applications to a common AMD software foundation.

MORE CHOICE FOR CUSTOMER WORKLOADS

AMD Instinct MI450X GPUs expand customer choice on AMD Helios with 288 GB of HBM4 per GPU. MI455X remains the 432 GB option for larger model footprints, longer contexts, or higher concurrency that benefit from added capacity. Both deliver the same core compute, bandwidth, and connectivity. Customers can select either or both configurations across their infrastructure. The best fit depends on model requirements, throughput and response-time targets, and overall deployment cost. MI450X brings another capacity choice to the same rack-scale architecture and ROCm software foundation.

LEARN MORE

For more information, visit AMD.com/INSTINCT.

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

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Instinct, CDNA, Infinity Fabric, ROCm, and combinations thereof are trademarks of Advanced Micro Devices, Inc. PCIe is a registered trademark of PCI-SIG Corporation. PyTorch, the PyTorch logo and any related marks are trademarks of The Linux Foundation. TensorFlow, the TensorFlow logo, and any related marks are trademarks of Google Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Use of third party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied GD-83