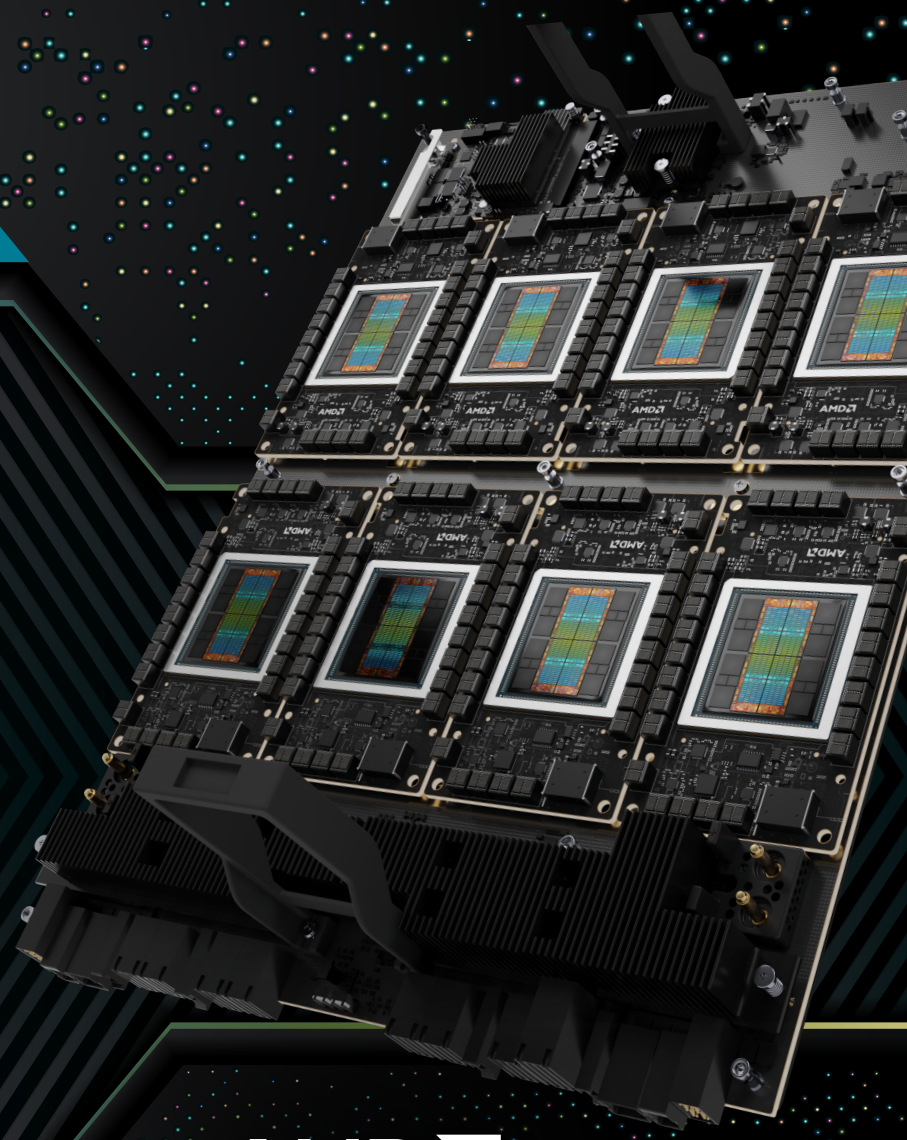
AMD INSTINCT

RAW ACCELERATION POWER FOR AI AND HPC

WITH AMD INSTINCT™ MI300X GPUs



AMD INSTINCT MI300X

As you tackle your greatest computing opportunities, it's time to pick up the pace. AMD Instinct™ MI300 Series accelerators and the ROCm™ 6.0 open-source software ecosystem help take AI and HPC to the next level.

AI IS IN PLAY

64% OF BUSINESS OWNERS anticipate AI will improve customer relationships and increase productivity.

60% OF BUSINESS OWNERS expect AI to drive sales growth.

53% OF BUSINESSES already apply AI to improve production processes, while

HIGH PERFORMANCE IS IN HIGH DEMAND

The global HPC market is expected to grow more than **6%** annually through 2027, to a total of more than **$51 BILLION**.

The expanded HPE-AI market will grow at **6.1%** to reach **$60.4 BILLION** in that time frame.[2]

51% OF BUSINESSES are adopting AI for process automation.[1]

GETTING THE MOST FROM THESE TECHNOLOGIES DEMANDS FOCUS ON:

SCALE

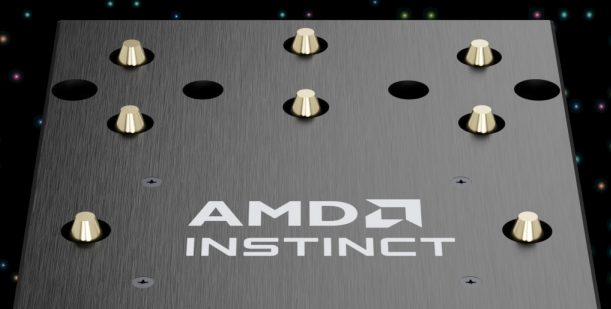
HPC systems have had to grow exponentially in processing capacity as the design, prediction and research models become larger and more complex. Meanwhile, dealing with the suddenly widespread interest in AI means building inference systems large enough to serve millions of users.

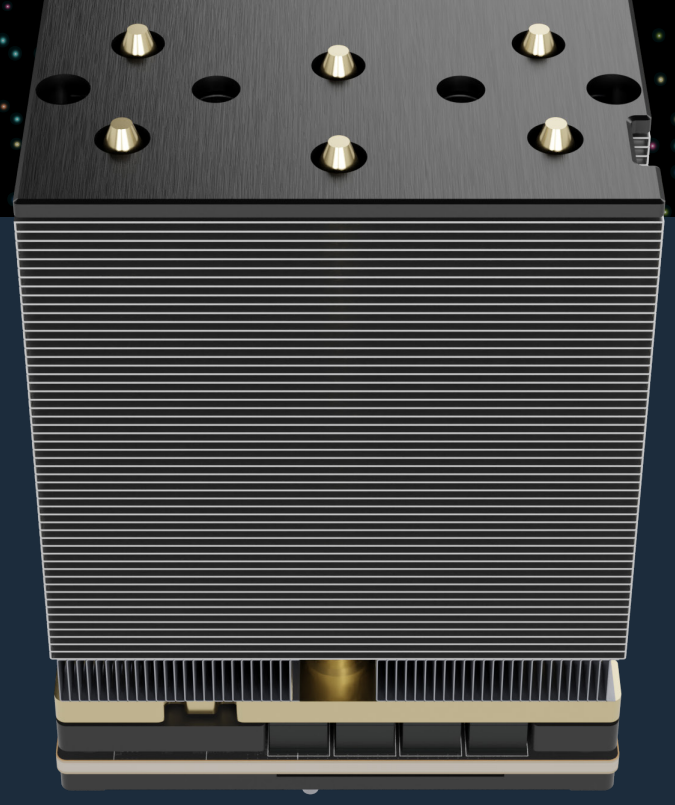
SPEED

First cloud and now AI have created a pervasive expectation in communities and markets of rapid availability of new technological capabilities and near-immediate results, but the scoping, design, construction, validation and deployment of high-powered, modern systems is is extremely difficult.

ARCHITECTURES

Critical AI and HPC systems need to be deployed close to existing data sources and dependent systems to avoid slowing down processing, processes and results. For IT organizations, it is a new advanced architecture to manage.





AMD INSTINCT

MI300X GPUs

LEADERSHIP *ACCELERATION* FOR AI AND HPC ON A *MULTI-GPU* PLATFORM

FASTER AI TRAINING AND INFERENCE

Superior performance in FP8, FP16, BF16, and TF32 calculations compared to the Nvidia H100 SXM5 (80GB).[MI300-18]

FASTER HPC JOBS

Superior performance in FP64 vector/matrix calculations compared to the Nvidia H100 SXM5 (80GB).[MI300-18]

FASTER DATA TRANSFERS

Up to 896GB/s peak aggregate theoretical GPU peer-to-peer I/O bandwidth performance in a ring of 8 AMD Instinct™ MI300X GPUs.[MI300-06]

MORE AI POWER

~2.5× the peak theoretical AI workload performance per watt (FP16) of previous-generation AMD Instinct GPUs.[MI300-27]

MORE *COMPUTE AND MEMORY* CAPACITY

GPUs AND MEMORY, PACKED

A fully connected 8-GPU platform offering an industry-leading 1.5TB of high-bandwidth memory (HBM3) capacity and 20.9 PFLOPS half-precision (FP16) and BFloat16 (BF16) peak theoretical floating-point performance with sparsity per node.[MI300-25, MI300-13]

HPC UPLIFT

3.4× the single-precision (FP32) peak theoretical floating-point performance of prior generation AMD Instinct GPUs for HPC.[MI300-11]

FEWER GPUs FOR TRAINING

The Hugging Face OPT transformer 66B-parameter large language model (LLM) can run on a single AMD Instinct MI300X GPU.[MI300-07]

EFFICIENT *POWER USAGE*

BREAKTHROUGH DESIGN

State-of-the-art die stacking and chiplet architecture delivers high density, enhanced integration of high-bandwidth memory and optimized energy consumption.

SPARSITY SUPPORT

For the first time on AMD accelerators, get hardware support for matrix sparsity, which helps save power, lower compute cycles and reduce memory use for some AI workloads.

EFFICIENT UTILIZATION

Spatial partitioning of CU inventory on each accelerator for use by more than one virtual client to help reduce capacity waste and increase hardware utilization.

MORE FROM EACH WATT

Up to a ~155% half-precision FP16 performance-per-watt gain compared to previous generation AMD Instinct MI250X 560W GPUs.[MI300-27]

SUSTAINABILITY STYLE

AMD Instinct powers 7 of the top 10 supercomputer systems on the Green500 list.[3]

HIGHLY *PROGRAMMABLE* AND *OPEN* GPU PLATFORM

AMD ROCm™ 6.0

Use the first open-source software platform for AI and HPC hyperscale-class GPU computing.

AMD ROCm

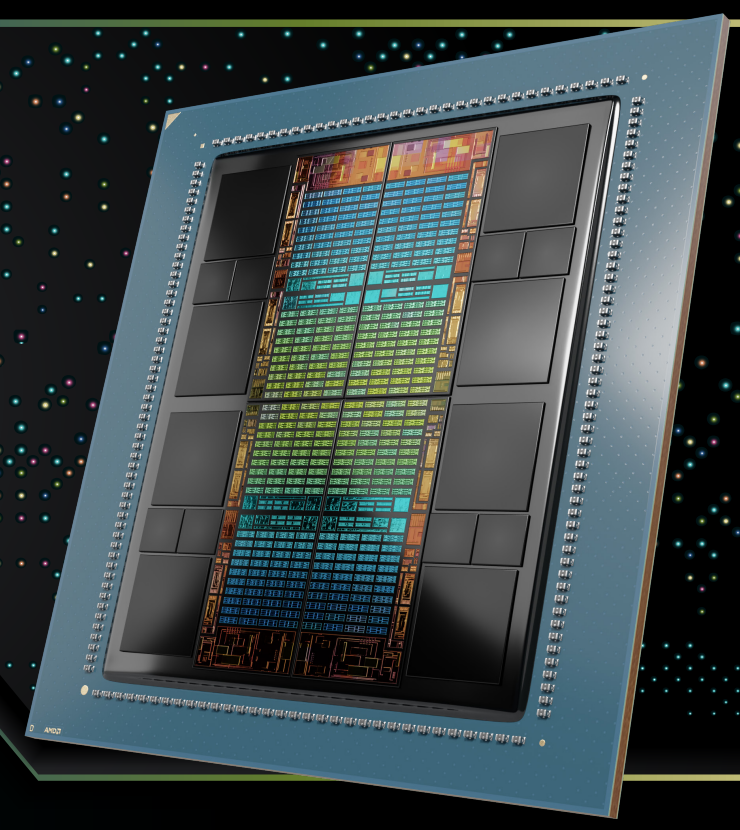
FRICTIONLESS ECOSYSTEM

Take advantage of drop-in support for major AI frameworks and HPC programming models, a key advantage for your evolving software needs.

POWERFUL TECHNICAL PARTNERSHIPS

AMD works with the Allen Institute for AI, Hugging Face, Lamini and OpenAI and continues to participate in open-source cross-platform initiatives such as MLIR, OpenMP®, OpenXLA, PyTorch, TensorFlow and Triton.

AMD INSTINCT



Power AI and HPC applications at scale with AMD Instinct™ MI300X accelerators.

Learn more at AMD.com/INSTINCT and AMD.com/ROCm

AMD
together we advance_data centers

1 Forbes Advisor, "How Businesses Are Using Artificial Intelligence In 2023," April 24, 2023, https://www.forbes.com/advisor/business/software/ai-in-business/
2 Intersect360 Research, "Worldwide HPC and AI 2022 Total Market Size and 2023–2027 Forecast: Products and Services, " May 16, 2023. https://www.intersect360.com/report/worldwide-hpc-and-ai-2022-total-market-size-and-2023-2027-forecast-products-and-services/
3 Top500, The Green500 list, November 2023, https://www.top500.org/lists/green500/2023/11/

©2023 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD arrow, AMD Instinct, Infinity Fabric, AMD CDNA, ROCm and combinations thereof, are trademarks of Advanced Micro Devices, Inc. OpenCL™ is a registered trademark used under license by Khronos. The OpenMP name and the OpenMP logo are registered trademarks of the OpenMP Architecture Review Board. TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc. PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc. NVIDIA is a trademark of NVIDIA Corporation in the U.S. and/or other countries. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

PID 232405395_I 12/23