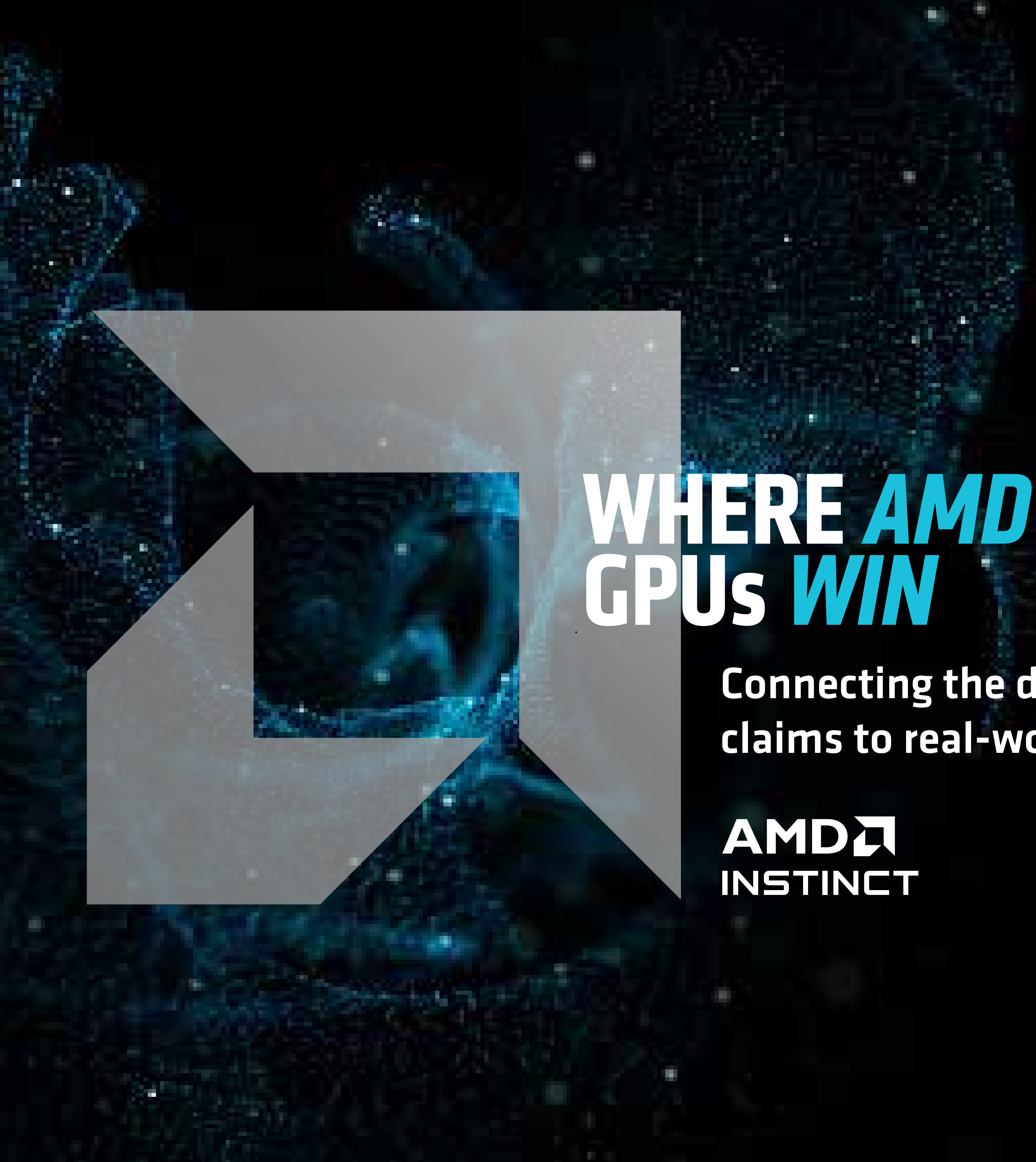


WHERE *AMD INSTINCT™* GPUs *WIN*

Connecting the dots from claims to real-world use cases

AMD INSTINCT

TABLE OF CONTENTS

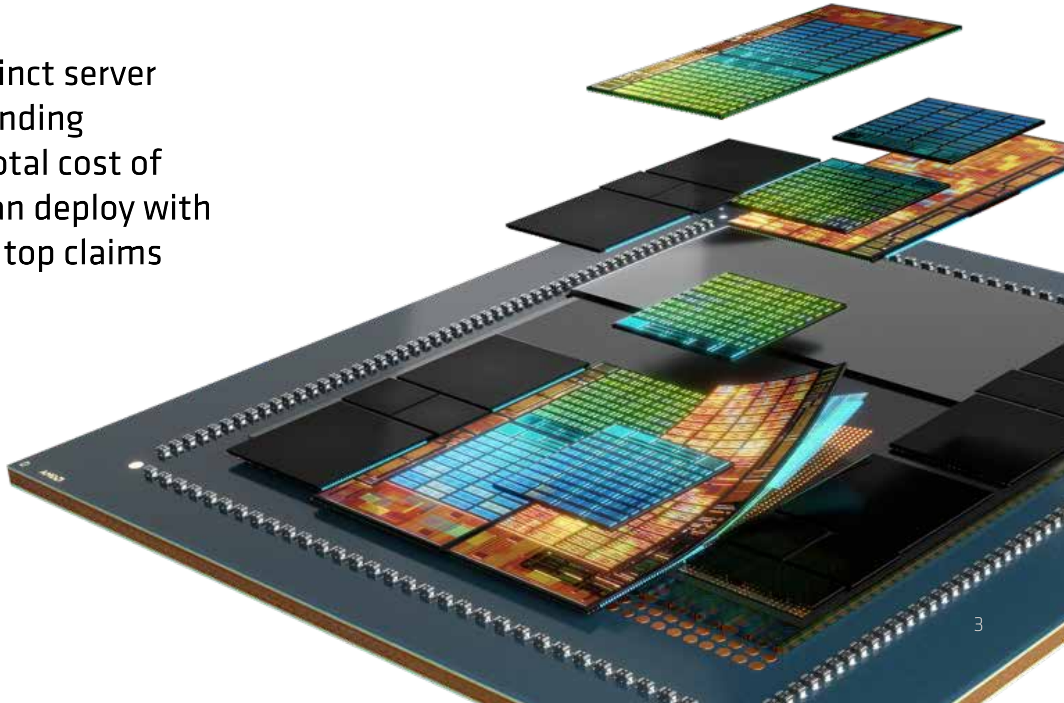
THE AI ERA DEMANDS CHOICE

The GPU market has long been dominated by a single vendor. But as the AI era gathers steam, the intense demand for data center server accelerators is highlighting the drawbacks of an unchallenged market.

Product shortages, premium pricing and extended lead times have created a demand for competition in the GPU market. Customers are actively searching for new solutions that can deliver better economics and an open ecosystem that allows vendor choice—without sacrificing performance.

CHOOSE PERFORMANCE WITH LOW TCO AND DEPLOY WITH CONFIDENCE

With AMD Instinct™ GPUs, AMD provides that choice. AMD Instinct server accelerators have leadership performance to supercharge demanding AI and HPC workloads. They give you more for less with lower total cost of ownership (TCO) and are designed for easy adoption – so you can deploy with confidence and without vendor lock-in. This ebook explores our top claims and how they impact results.



PERFORMANCE CLAIMS TO REAL-WORLD USE CASES

LARGER MEMORY CAPACITY
BY THE NUMBERS

Up to
256 GB
per AMD Instinct™ MI325X GPU

Up to
192 GB
per AMD Instinct MI300X GPU

Up to
128 GB
per AMD Instinct MI250 GPU

Up to
80 GB[1]
per NVIDIA A100 Tensor Core GPU

REAL-WORLD IMPACT

BETTER PERFORMANCE

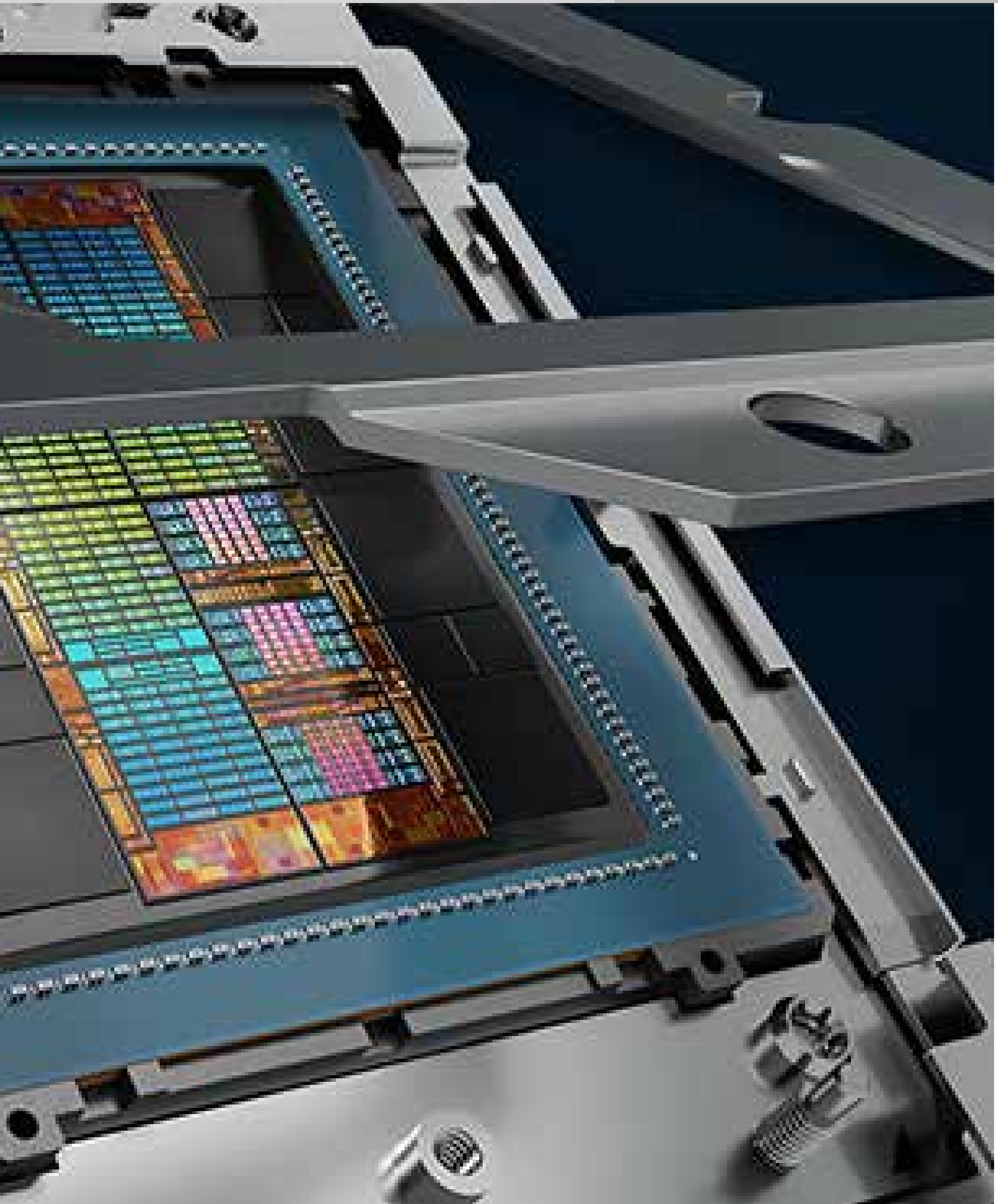
Eliminates memory bottlenecks and enhances processing speed and accuracy

LARGER WORKLOADS

Supports large models or datasets and high-batch processing, ideal for larger AI models and HPC workloads

FASTER TRAINING AND INFERENCING

Enables loading complex data structures or large models efficiently in low-batch settings



USE CASES THAT BENEFIT

DEEP LEARNING

Large neural networks used in deep learning require extensive amounts of memory for both training and inferencing. Larger GPU memory allows for handling more complex models and larger datasets, such as with large language models (LLMs).

SCIENTIFIC SIMULATIONS

Climate modeling, astrophysics, molecular dynamics and other scientific simulations require processing vast amounts of data. Increased GPU memory capacity enables systems to handle larger and more detailed simulations.

GENOMICS AND BIOINFORMATICS

DNA sequencing and other bioinformatics tasks require analyzing massive datasets, often involving complex algorithms. More GPU memory enables researchers to handle larger genomes and perform analyses more efficiently.

COMPUTER VISION

Applications like autonomous vehicles, robotics and surveillance systems rely on computer vision algorithms that process high-resolution images and videos. Larger GPU memory helps improve the performance of these applications.

FASTER MEMORY BANDWIDTH BY THE NUMBERS

Up to
6 TB/s
per AMD Instinct™ MI325X GPU

Up to
5.2 TB/s
per AMD Instinct MI300X GPU

Up to
3.2 TB/s
per AMD Instinct MI250 GPU

Up to
2 TB/s[1]
per NVIDIA A100 Tensor Core GPU

REAL-WORLD IMPACT

INCREASED PERFORMANCE

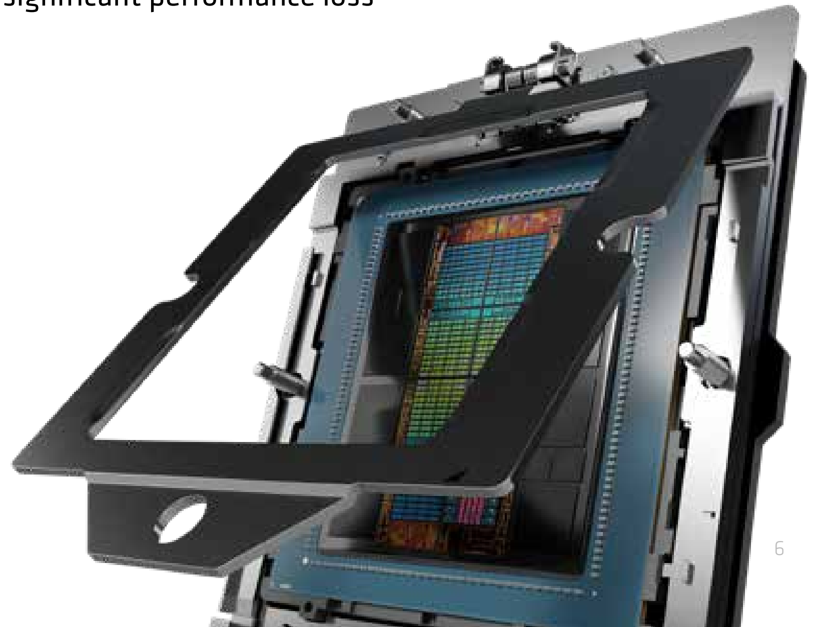
Reduces latency for data transfer and processing, resulting in improved overall performance for AI and HPC workloads

REDUCED BOTTLENECKS

Helps alleviate bottlenecks in data-intensive tasks, leading to smoother and more efficient performance

BETTER MULTITASKING

Allows the GPU to handle multiple tasks simultaneously without significant performance loss



USE CASES THAT BENEFIT

DEEP LEARNING

Training and inferencing of deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), rely heavily on fast memory access for efficient processing of large datasets.

SCIENTIFIC SIMULATIONS

Large-scale simulations in fields like climate modeling, molecular dynamics and astrophysics require rapid data processing and analysis, which can be significantly improved with faster memory bandwidth.

GENOMICS AND BIOINFORMATICS

Tasks such as language translation, speech recognition and sentiment analysis involve processing large amounts of text data, making faster memory bandwidth essential for efficient processing.

COMPUTER VISION

Faster memory bandwidth allows for quick processing of large amounts of historical and real-time financial data, running complex algorithms, and performing Monte Carlo simulations to calculate risk and predict market trends.

HIGHER PERFORMANCE FOR LOW-BATCH AI PROCESSING

BY THE NUMBERS

MEMORY CAPACITY

Up to

256 GB

per AMD Instinct™ MI325X GPU

Up to

80 GB

per NVIDIA A100 Tensor Core GPU

MEMORY BANDWIDTH

Up to

6 TB/s

per AMD Instinct MI325X GPU

Up to

2 TB/s

per NVIDIA A100 Tensor Core GPU

REAL-WORLD IMPACT

EXCELLENT PERFORMANCE

Better memory capacity combined with higher memory bandwidth leads to excellent performance for workloads with small to medium batch sizes

REAL-TIME RESPONSIVENESS

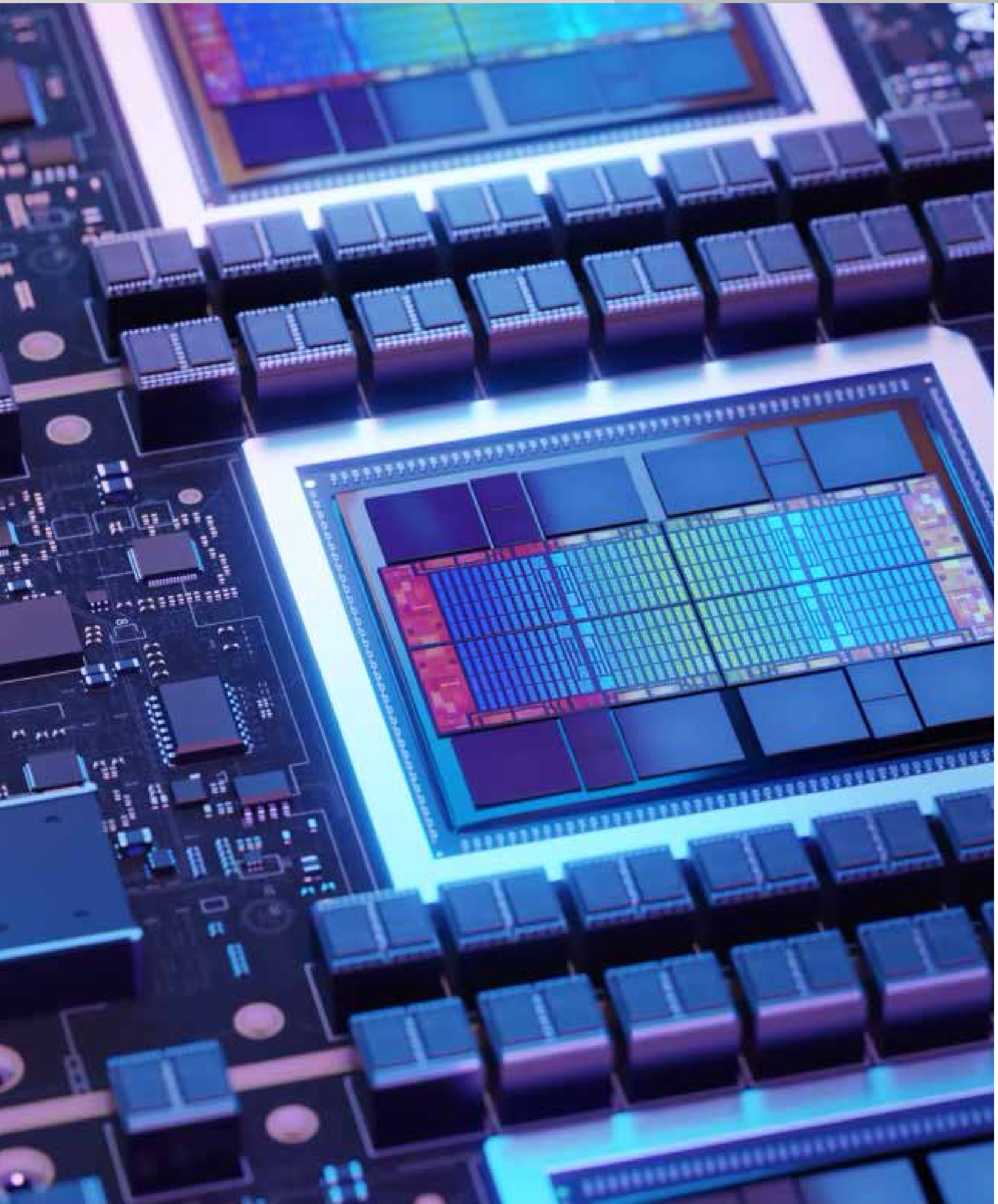
Minimizes latency for quick request processing without memory swaps or delays, making it ideal for real-time applications that rely on low-batch processing

HIGHER EFFICIENCY IN SPARSE DATA ENVIRONMENTS

Supports extensive model operations with real-time analysis by enabling efficient processing of intermittent data without waiting for larger batches

SUPPORT FOR HYBRID WORKLOADS

Grants the flexibility to switch between high-batch and low-batch tasks, handling real-time, single-inference tasks and batch-processing workloads without reconfiguring memory



USE CASES THAT BENEFIT

AI TRAINING FOR IMAGE RECOGNITION

Training and deploying image recognition models often require processing small batches of images in real time to ensure accurate object detection and classification.

SCIENTIFIC RESEARCH

Many scientific simulations and analyses involve iterative processing of small batches of data so researchers can explore different scenarios and fine-tune models.

REAL-TIME CONVERSATIONAL AI

Virtual assistants and chatbots rely on small to medium batch sizes to process natural language queries and generate responses in real time.

AUTONOMOUS VEHICLES

Self-driving vehicles need to process small to medium batches of sensor data in real time to make timely decisions and navigate safely through complex environments.

HEALTHCARE MONITORING

Wearable devices and remote patient monitoring systems rely on processing small batch data in real time to detect anomalies, provide alerts and facilitate timely interventions.

INDUSTRY-LEADING FP64 PERFORMANCE

BY THE NUMBERS

FP64 PERFORMANCE

Up to

2.4X faster

compared to the NVIDIA H200 SXM MI325-008

FP64 (VECTOR)

81.7 TFLOPS

per AMD Instinct™ MI325X GPU

33.5 TFLOPS

compared to the NVIDIA H200 SXM[1]

FP64 (TENSOR/MATRIX)

163.4 TFLOPS

per AMD Instinct ML325X GPU

66.9 TFLOPS

compared to NVIDIA H200 SXM[2]

REAL-WORLD IMPACT

ENHANCED ACCURACY

Double-precision (FP64) arithmetic provides higher precision and accuracy compared to single-precision (FP32) or mixed-precision formats, leading to more reliable results in simulations, modeling and data analysis.

FASTER COMPUTATION

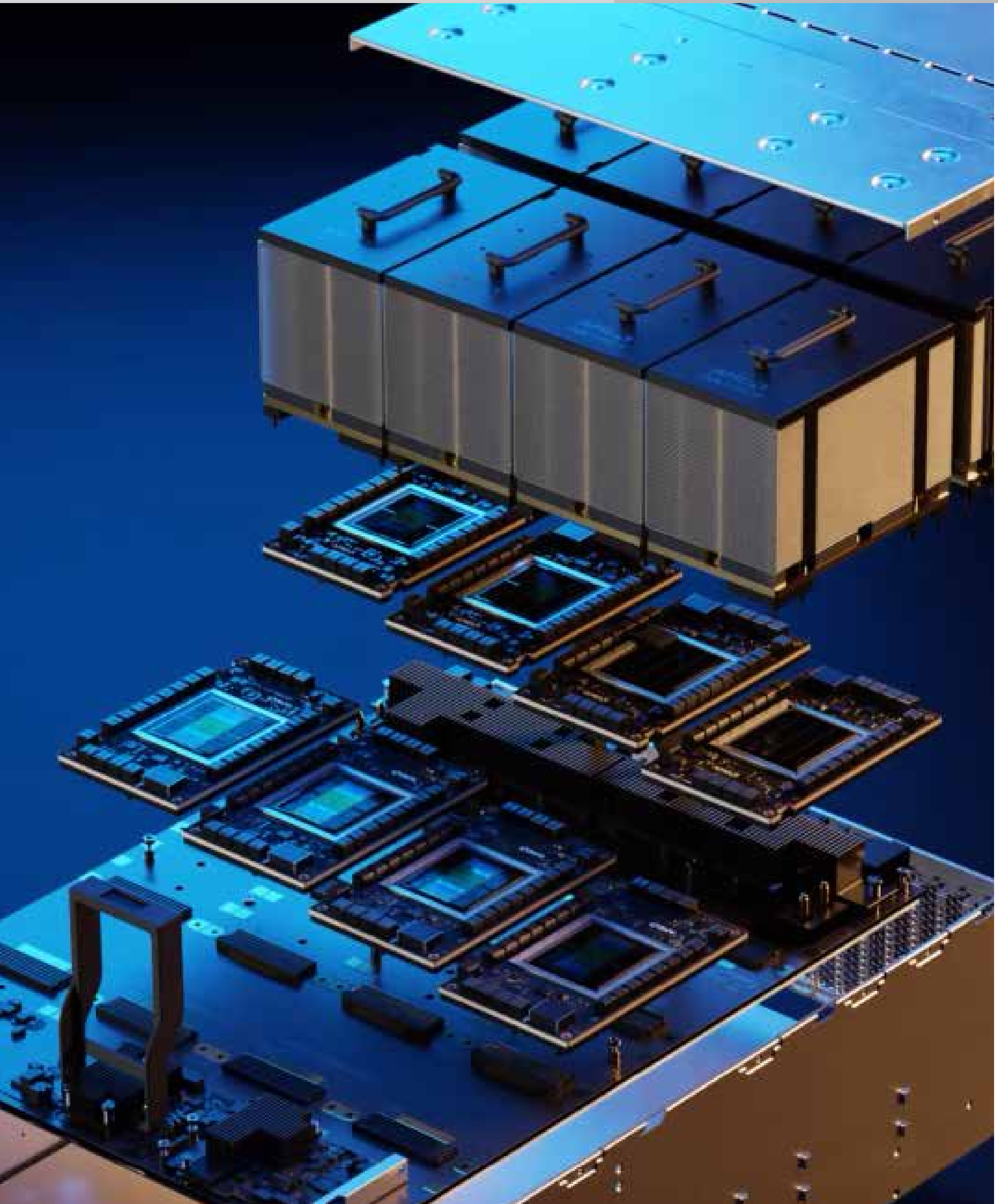
Accelerates the overall computation time for HPC and scientific computing applications.

IMPROVED COMPATIBILITY

Strong FP64 performance leads to better compatibility with existing scientific computing codes and libraries that rely on double-precision arithmetic, facilitating smoother migration to GPU-accelerated platforms.

WIDER APPLICATION RANGE

Expands the range of applications that can benefit from GPU acceleration to include computational fluid dynamics, structural analysis, quantum chemistry simulations and others.



USE CASES THAT BENEFIT

FINANCIAL MODELING

FP64 performance is needed for financial simulations and risk assessments as well as trading algorithms that require high precision.

ENGINEERING SIMULATIONS

Aerospace, automotive and civil engineering applications—including finite element analysis and computational fluid dynamics—benefit from high-precision FP46 performance.

ASTROPHYSICS AND SPACE EXPLORATION

Simulating celestial bodies and space missions and modeling the universe's structure require the highest levels of precision.

WEATHER AND CLIMATE FORECASTING

Detailed atmospheric process simulations and long-term climate predictions and extreme weather modeling benefit from FP64 performance.

QUANTUM MECHANICS AND PARTICLE PHYSICS

Managing the complexities of quantum systems and simulating particle interactions rely on highly precise calculations.

SEISMIC AND GEOPHYSICAL ANALYSIS

Enhanced accuracy benefits the analysis of seismic data and modeling of Earth's subsurface structures.

BETTER COST EFFICIENCY, ENERGY EFFICIENCY AND SUSTAINABILITY

BY THE NUMBERS

- **6X performance**
  within a comparable power envelope[3]
- **1.2X higher efficiency**
  compared to the NVIDIA H100 SXM MI300-25
- **Up to ~2.5X**
  peak theoretical generative AI and training workload FP16 performance per watt with CDNA™ architecture advancements MI300-27
- **2X higher**
  performance per watt on HPL-AI and HPL benchmarks MI200-81, MI200-69A
- **3 of the top 10**
  systems on the Green500 list are powered by AMD Instinct™ accelerators[4]
- **1.2X higher efficiency**
  than the NVIDIA H100 SXM for consolidation savings MI300-25

REAL-WORLD IMPACT

LOWER OPERATIONAL COSTS

Optimizes the use of space, power and budget resources

SUSTAINABLE AI AND HPC

Helps reduce the environmental impact of AI and sustained processing in research labs and HPC centers

POWER SAVINGS AT SCALE

Lowers power consumption to reduce operating costs and environmental impact for large workloads, including demanding scientific and AI applications

ECO-FRIENDLY HIGH BANDWIDTH

Minimizes resource usage by accelerating data access to drive faster results while conserving data center resources

USE CASES THAT BENEFIT

LARGE-SCALE AI TRAINING AND INFERENCING

More power-efficient GPUs can significantly reduce data center energy consumption and operating costs.

SCIENTIFIC RESEARCH

Reduce the carbon footprint of scientific research projects that require extensive HPC resources for simulations, data analysis and modeling.

EDGE

Edge devices like autonomous vehicles, drones and IoT gateways rely on real-time AI processing while operating within tight power and thermal constraints.

MOBILE DEVICES

Power-efficient GPUs are crucial for smartphones, tablets and other mobile devices that use AI for various applications like image processing, natural language processing and augmented reality.

RENEWABLE ENERGY SYSTEMS

Optimize the performance and energy consumption of renewable energy systems that rely on AI and HPC for weather prediction, grid optimization and energy storage management.

LOWER LATENCY AT SCALE

BY THE NUMBERS

- **#1 and #2 supercomputers**
  El Capitan and Frontier on the Top500 list are powered by AMD.[5]
- **#1 fastest commercial system and #1 fastest in Europe**
  ENI's HPC6 system on the Top500 list is powered by AMD.[5]
- **5 of the top 10 fastest**
  on the Top500 list are powered by AMD.[5]
- **22 of the top 50 fastest**
  on the Top500 list are powered by AMD.[5]
- **15 of the most energy-efficient**
  on the Green500 list are powered by AMD.[4]

REAL-WORLD IMPACT

CONSISTENT PERFORMANCE AT LARGE BATCH SIZES

Better latency management provides consistent performance even when handling large batch sizes, making it ideal for HPC workloads that demand high throughput and stable execution times.

STABLE PERFORMANCE UNDER PEAK LOADS

Consistent performance even when the GPU is operating at peak capacity prevents slowdowns and provides reliable results.

FASTER TRAINING AND INFERENCING

Accelerating the execution of AI training and inferencing tasks results in faster model development, deployment and decision-making.

HIGH THROUGHPUT

GPUs can process larger volumes of data in a given time frame, enhancing overall system throughput and productivity.

MINIMIZED PROCESSING LAG

Helps deliver smoother user experiences and more responsive real-time applications, such as autonomous systems and interactive data visualization.

USE CASES THAT BENEFIT

FINANCIAL SERVICES

Low latency is crucial for real-time fraud detection, algorithmic trading and risk analysis, where even milliseconds can significantly impact outcomes.

HEALTHCARE MONITORING

In applications like telemedicine and remote patient monitoring, low latency enables timely processing and analysis of patient data, enabling swift medical interventions when needed.

IMAGE RECOGNITION

Faster, low-latency processing accelerates the training and deployment of image recognition models for applications such as autonomous vehicles, security systems and manufacturing quality control.

SCIENTIFIC RESEARCH

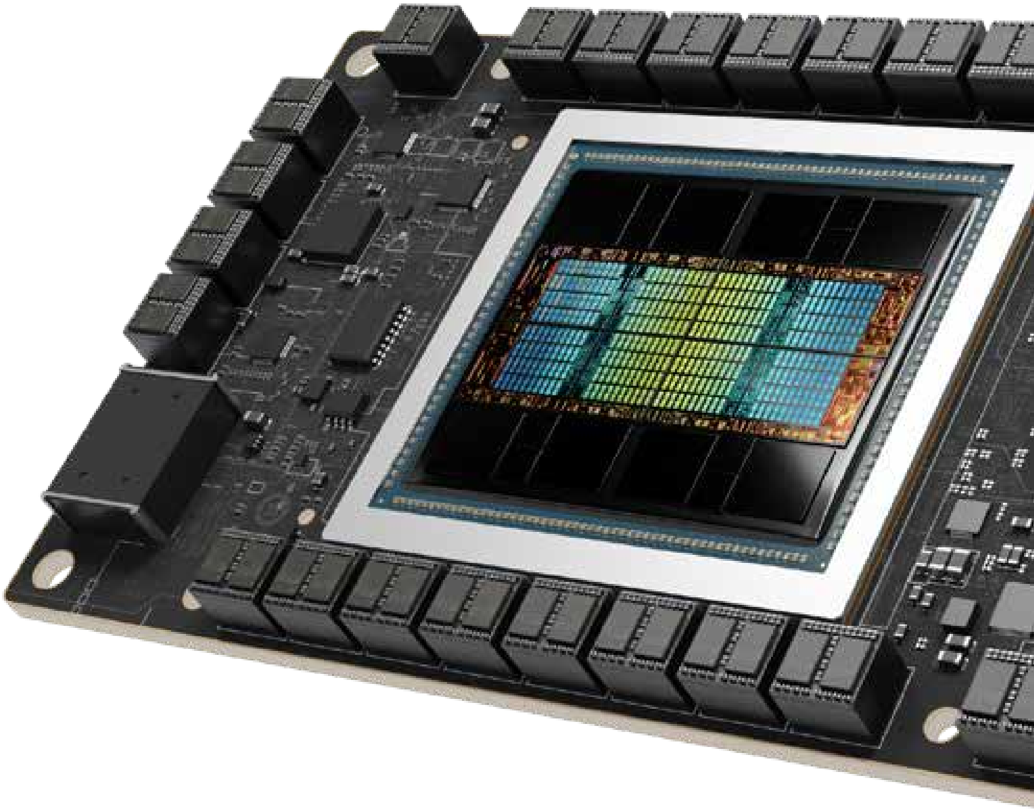
Lower latency enables faster processing of massive datasets, leading to quicker insights and accelerated discovery for use cases such as climate modeling, weather forecasting and computational fluid dynamics.

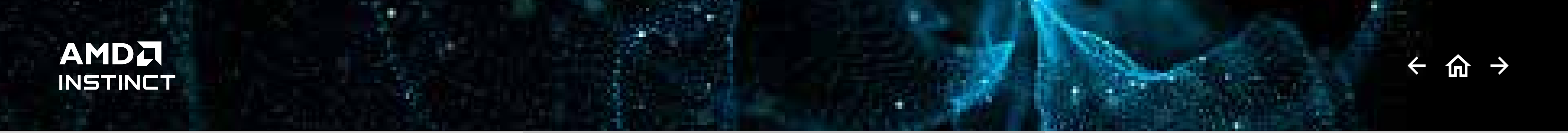
NATURAL LANGUAGE PROCESSING

Low latency is vital for real-time applications like conversational AI, language translation and sentiment analysis.

EDGE COMPUTING

Lower latency facilitates efficient processing and decision-making for real-time edge applications such as robotics, IoT and augmented reality.





OPEN AND EXPANDING ECOSYSTEM BY THE NUMBERS

1,000,000+ models
supported on Hugging Face

Day-zero support
for industry-leading AI frameworks and HPC programming models such as PyTorch, JAX and Triton

Tier 1 and Tier 2
cloud service provider (CSP) partnerships including Microsoft, Meta, Oracle Cloud, Vultr, NSCALE and others

Deep co-development
with some of the largest global technology leaders, including the Allen Institute for AI, Hugging Face, Lamini and OpenAI

REAL-WORLD IMPACT

OPEN SOURCE

AMD Instinct™ GPUs and the open-source AMD ROCm™ software platform provide easy-to-use tools that are built around industry standards and enable developers to create well-optimized portable software for AI and HPC.

SUPPORT FOR HYBRID, MULTICLOUD ENVIRONMENTS

Open portability supports environments across multiple vendors and architectures.

VENDOR INDEPENDENCE

Facilitates adoption of multiple acceleration platforms and cross-platform HPC and AI development for greater flexibility and choice in hardware selection.

OPTIMIZED RESOURCE USAGE

Customizable hardware setups can be easily optimized for specific needs.

FRICTIONLESS SOFTWARE ECOSYSTEM

Includes drop-in support for major AI and HPC frameworks, leading software platforms and models, and networking infrastructures to help ease system deployment.

USE CASES THAT BENEFIT

AI AND HPC IN THE CLOUD

Enables use of cost-effective, flexible, scalable cloud-based infrastructures with interoperability that transcends vendor lock-in and enables switching between CSPs or adopting new technologies without significant disruption or costs.

COLLABORATIVE HPC

Open-source GPU ecosystems facilitate collaboration among researchers and developers, fostering innovation and enabling the co-development of novel HPC and AI solutions across domains.

MULTINODE DISTRIBUTED COMPUTING

Open-source GPU solutions enable users to customize and optimize multinode distributed computing environments according to specific requirements, for better performance and resource utilization.

EDGE COMPUTING

Open-source software is often more customizable and interoperable, making it easier to integrate with various edge computing platforms, frameworks and tools to facilitate development of versatile and adaptable edge computing solutions.

AMD INSTINCT™ GPUS: THE CHOICE YOU'VE BEEN WAITING FOR

AMD Instinct™ accelerators supercharge AI and HPC to help you get new insights and make new discoveries. Choose from a wide portfolio to best meet your needs.

AMD INSTINCT™ MI300 SERIES ACCELERATORS

- Leadership performance for the data center at any scale, from single-server solutions to exascale-class supercomputers.
- Uniquely well-suited to power demanding AI and HPC workloads with exceptional performance, large memory density, high bandwidth memory and support for specialized data formats.

AMD INSTINCT™ MI325X ACCELERATOR

Leadership memory capacity and bandwidth are ideal for GenAI.

AMD INSTINCT MI300X PLATFORM

Eight AMD Instinct MI300X accelerators in a single UBB form factor interconnected to maximize compute performance.

AMD INSTINCT MI300A APU

AMD Instinct and AMD EPYC™ processors are combined with shared memory for enhanced flexibility, efficiency and management.

AMD INSTINCT MI200 SERIES ACCELERATORS

- A powerful platform to fuel the convergence of HPC and AI
- Delivers a quantum leap in HPC and AI performance over competitive GPUs
- Up to 4X advantage in HPC performance of the Instinct MI250X GPU over the NVIDIA A100 GPU [MI200-01]
- The first data center GPU to deliver 383 teraFLOPS of theoretical mixed-precision FP16 performance for deep learning training [MI200-01]

TRUST YOUR INSTINCTS

AMD.com/Instinct

[1] NVIDIA, "NVIDIA A100 Tensor Core GPU," accessed November 2024.
[2] AMD, "AMD Instinct™ MI325X Accelerators: HPC Performance," accessed November 2024.
[3] AMD, "Achieving 6x the Performance with HPE and AMD," accessed November 2024.
[4] Top500, "Green500 list," November 2024.
[5] Top500, "Top500 list," November 2024.

DISCLAIMER

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u

Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

END NOTES

For details on the claims used in this document, visit amd.com/en/legal/claims/instinct.

MI200-01
World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance.
Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance.
Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison.
https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.

MI200-69A
Testing Conducted by AMD performance lab 11/14/2022 using HPL comparing two systems. 2P EPYC™ 7763 powered server, SMT disabled, with 1x, 2x, and 4x AMD Instinct™ MI250 (128 GB HBM2e) 560W GPUs, host ROCm 5.2.0 rocHPL6.0.0. AMD HPL container is not yet available on Infinity Hub. vs. 2P AMD EPYC™ 7742 server, SMT enabled, with 1x, 2x, and4x Nvidia Ampere A100 80GB SXM 400W GPUs, CUDA 11.6 and Driver Version 510.47.03. HPL Container (nvcr.io/nvidia/hpc-benchmarks:21.4-hpl) obtained from https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI200-81
HPL-AI comparison based on AMD internal testing as of 11/2/2022 measuring the HPL-AI benchmark performance (TFLOPS) using a server with 2x EPYC™ 7763 with 4x MI250 (128MB HBM2e) with Infinity Fabric running host ROCm™ 5.2.0, HPL-AI-AMD v1.0.0; AMD HPL-AI container not yet available on Infinity Hub.
Versus a server with 2x EPYC 7742 with 4x A100 SXM (80GB HBM2e) running CUDA® 11.6, HPL-AI-NVIDIA v2.0.0, container nvcr.io/nvidia/hpc-benchmarks:21.4-hpl.
Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-25
Measurements conducted by AMD Performance Labs as of November 18th, 2023 on the AMD Instinct™ MI300X (192 GB HBM3) 750W GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 1307.4 TFLOPS peak theoretical half precision (FP16), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16).
The MI300X is expected to be able to take advantage of fine-grained structure sparsity providing an estimated 2x improvement in math efficiency resulting 2,614.9 TFLOPS peak theoretical half precision (FP16), 2,614.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 floating-point performance with sparsity.
Published results on Nvidia H100 SXM (80GB HBM3) 700W GPU resulted in 989.4 TFLOPS peak theoretical half precision (FP16 Tensor), 989.4 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor), 1,978.9 TFLOPS peak theoretical half precision (FP16 Tensor) with sparsity, 1,978.9 TFLOPS peak theoretical Bfloat16 format precision (BF16 Tensor) with sparsity floating-point performance. Nvidia H100 source: https://resources.nvidia.com/en-us-tensor-core/
AMD Instinct™ MI300X AMD CDNA 3 technology-based accelerators include up to eight AMD Infinity Fabric links providing up to 1,024 GB/s peak aggregate theoretical GPU peer-to-peer (P2P) transport rate bandwidth performance per GPU OAM module.

MI300-27
Calculations conducted by AMD Performance Labs as of Nov 29, 2023, for the AMD Instinct™ MI300X (192GB HBM3 OAM Module) 750W accelerator designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 1,307.44 TFLOPS peak theoretical half precision (FP16) floating-point performance.
The AMD Instinct™ MI300A (128GB HBM3 APU) 760W accelerator designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 980.58 TFLOPS peak theoretical half precision (FP16) floating-point performance.
The AMD Instinct™ MI250X (128GB HBM2e OAM module) 560W accelerator designed with AMD CDNA™ 2 6nm FinFET process technology at 1,700 MHz peak boost engine clock resulted in 383.0 TFLOPS peak theoretical half precision (FP16) floating-point performance.

MI325-008
Calculations conducted by AMD Performance Labs as of October 2nd, 2024 for the AMD Instinct™ MI325X (1000W) GPU designed with AMD CDNA™ 3 5nm | 6nm FinFET process technology at 2,100 MHz peak boost engine clock resulted in 163.4 TFLOPs peak theoretical double precision Matrix (FP64 Matrix), 81.7 TFLOPs peak theoretical double precision (FP64), 163.4 TFLOPs peak theoretical single precision Matrix (FP32 Matrix), 163.4 TFLOPs peak theoretical single precision (FP32), 653.7 TFLOPS peak theoretical TensorFloat-32 (TF32), 1307.4 TFLOPS peak theoretical half precision (FP16), Actual performance may vary based on final specifications and system configuration.
Published results on Nvidia H200 SXM (141GB) GPU: 66.9 TFLOPs peak theoretical double precision tensor (FP64 Tensor), 33.5 TFLOPs peak theoretical double precision (FP64), 66.9 TFLOPs peak theoretical single precision (FP32), 494.7 TFLOPs peak TensorFloat-32 (TF32), 989.5 TFLOPS peak theoretical half precision tensor (FP16 Tensor). TF32 Tensor Core performance were published by Nvidia using sparsity; for the purposes of comparison, AMD converted these numbers to non-sparsity/dense by dividing by 2, and this number appears above.
Nvidia H200 source: https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23-h200-datasheet-3002446 and https://www.anandtech.com/show/21136/nvidia-at-sc23-h200-accelerator-with-hbm3e-and-jupiter-supercomputer-for-2024
Note: Nvidia H200 GPUs have the same published FLOPs performance as H100 products https://resources.nvidia.com/en-us-tensor-core/.
* Nvidia H200 GPUs don't support FP32 Tensor.