

Reconciling AMD Computing Power with Energy Efficiency in AI Factories

Forum Teratec 2025
May 22nd, 2025

Jose Noudohouenou

AMD 
together we advance_

AMD PLATFORM FOR ACCELERATED COMPUTING

INVESTING IN CORE TECHNOLOGY FOR LEADERSHIP IN HPC & AI

COMPUTE

WORKLOAD-OPTIMIZED
COMPUTE ARCHITECTURE
w/ WIDE RANGE OF DATA
FORMAT SUPPORT

AMD
CDNA

FP64 FP32

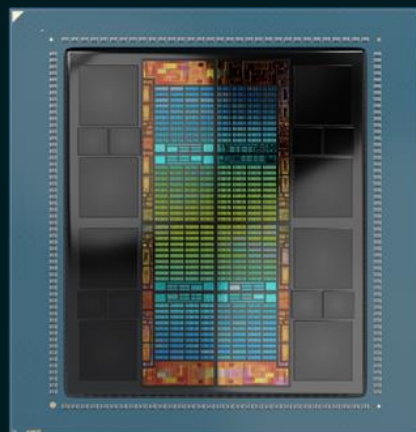
FP16 BF16

FP8 INT8

And more...

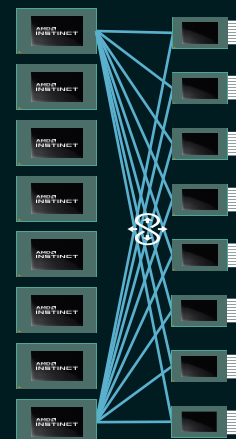
MEMORY

HIGHEST CAPACITY
MEMORY AND BANDWIDTH
AVAILABLE IN THE
INDUSTRY



NETWORKING

ADVANCING NETWORK
BANDWIDTH WITH
SUPPORT FOR INDUSTRY
STANDARD & CUSTOM
TECHNOLOGIES



SOFTWARE

FRICITIONLESS SW
ECOSYSTEM W/DROP-IN
SUPPORT FOR LEADING
PROGRAMMING MODELS &
AI FRAMEWORKS

OpenMP

kokkos

RAJA

PyTorch

TensorFlow

deepspeed

Triton

OpenXLA

Powering the Top 2 Supercomputers in the world

2 Generations of Instinct™ - MI300A & MI250X leading Top500



5 of Top10



15 of Top25

#1 EL CAPITAN

#2 FRONTIER

#5 eni

#8 LUMI

#10 Lawrence Livermore National Laboratory

GENCI

INES 69 GF/W

Lawrence Livermore National Laboratory 63 GF/W

FRONTIER 63 GF/W

Top 10

4.2_{EF} | 61%
AMD Instinct™

1.2_{EF} | 18%
Nvidia

AMD Instinct™ MI300A 69_{GF/W}

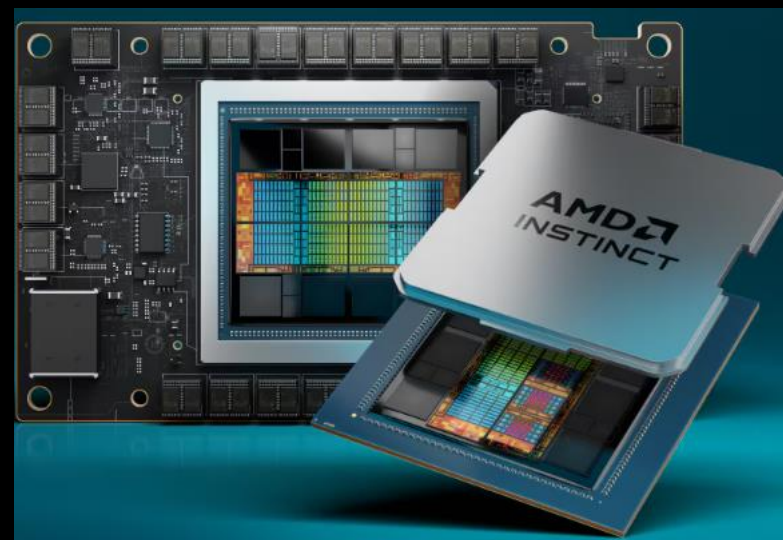
AMD Instinct™ MI250X 63_{GF/W}

AMD HPC & AI Computing Resources Overview



AMD HPC & AI Computing Resources Portfolio

- AMD product portfolio
 - addressing the widest set of customer needs
 - enabling the most AI use-cases from the cloud to the edge to endpoints.
- Product List:
 - AMD EPYC™ Server processors
 - AMD Instinct™ Accelerators
 - AMD Adaptive SoCs
 - AMD Pensando™ DPU Accelerators
 - AMD Ryzen™ Processors
 - AMD Alveo™ Adaptive Accelerators
- Key products for AI Factories data centers:
 - AMD EPYC™ Server processors (CPU)
 - AMD Instinct™ Accelerators (GPU)
 - AMD Pensando™ DPU Accelerators (Networking)



Some AMD Instinct™ Accelerators



Launch Year	Accelerator	# Computing Units	Memory (HBM)	Power(W)
2024	MI325X	304	256	1000
2023	MI300X	304	192	750
2023	MI300A (APU)	228	128	760 (CPU+GPU)
2021	MI250X	220	128	560
2021	MI250	208	128	560

Thermal: Liquid Cooled & Air Cooled

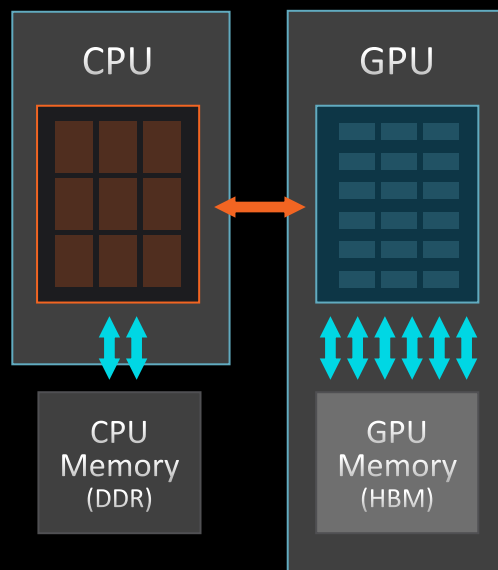
- Over the years:
- More compute units
 - More memory (and BW)
 - More power-hungry GPUs
 - New datatypes support at HW level
(well suited for AI applications/benchmarks)

https://en.wikipedia.org/wiki/AMD_Instinct

AMD MI300A

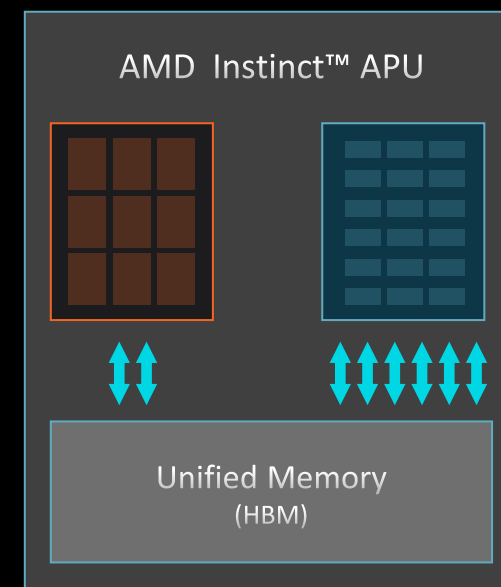
UNIFIED MEMORY APU ARCHITECTURE BENEFITS

AMD CDNA™ 2 Coherent Memory Architecture



AMD CDNA™ 3 Unified Memory APU Architecture

- Eliminate Redundant Memory Copies
- No programming distinction between host and device memory spaces
- High performance, fine-grained sharing between CPU and GPU processing elements
- Single process can address all memory, compute elements on a socket



Available 2025

AMD Instinct™ MI350 Series

Continued Gen AI Leadership

AMD
CDNA 4

3nm
Process Node

Up to
288GB
HBM3E

NEW
FP4 / FP6
Datatype Support

Efficient AMD Instinct™ Accelerators Utilization



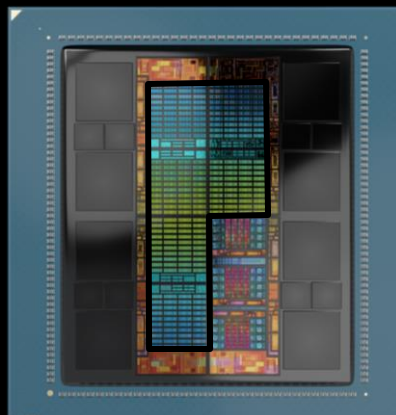
ENABLING MULTIPLE WORKLOADS FOR OPTIMAL GPU UTILIZATION

MI300 PARTITIONING

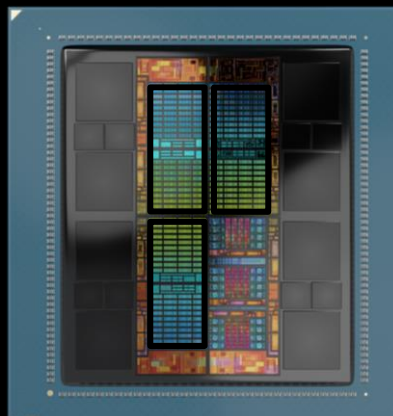
MI300A - APU

- Maximize GPU utilization with 3 partitions
- NPS Modes* (NPS1)

Single partition



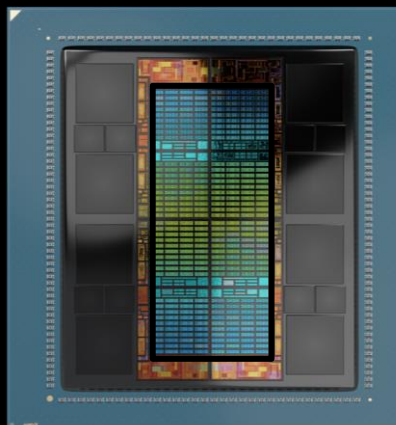
Three partitions



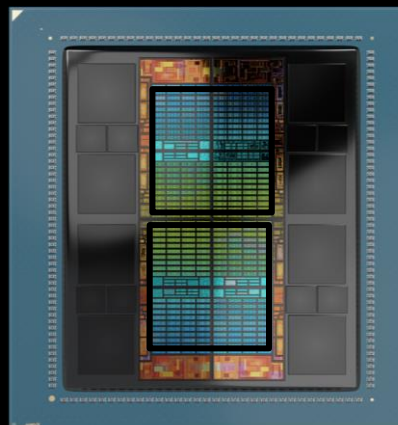
MI300X - OAM

- Maximize GPU utilization with up to 8 partitions
- NPS modes* (NPS1, NPS4)

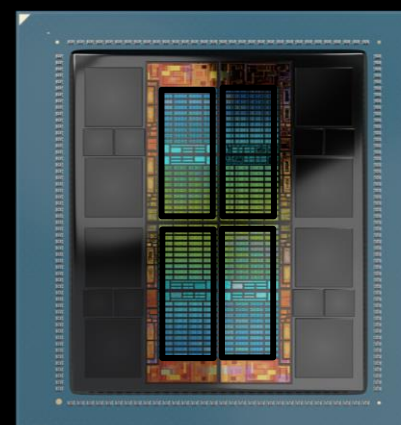
Single partition



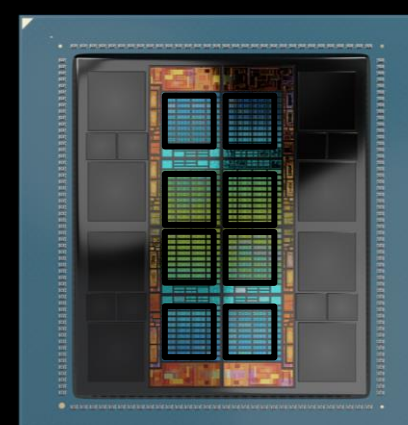
Two partitions



Four partitions



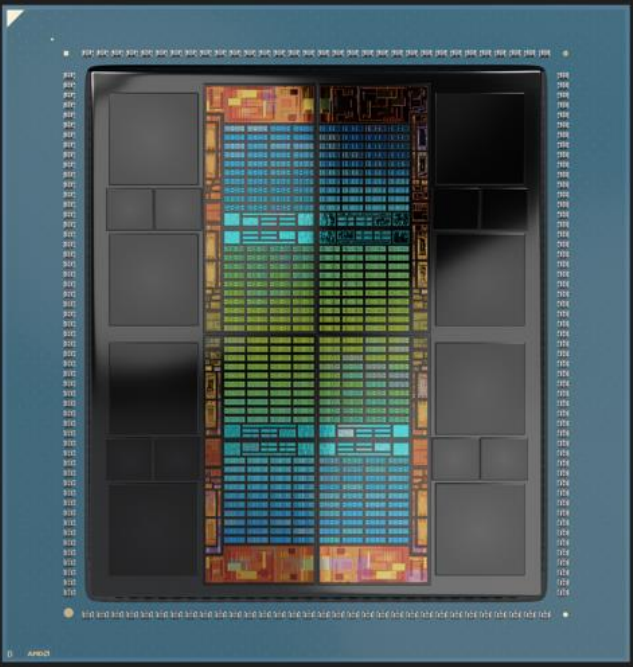
Eight partitions



* → Memory partitioning can only be changed via a re-boot

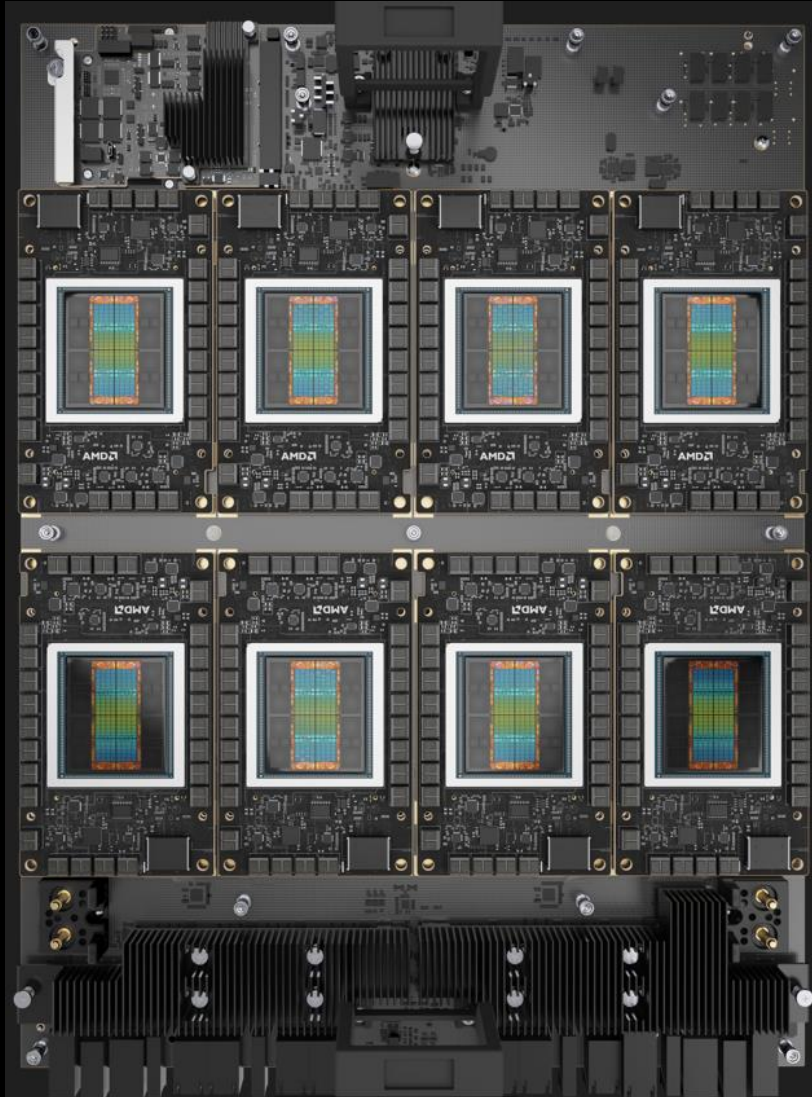
AMD Instinct™ MI300X GPU Partitioning

Support for up to 8 SR-IOV Virtual Functions per GPU

GPU Partition Options	1 (SPX)	2 (DPX)	4 (QPX)	8 (CPX)
	<div>GPU Instance 192GB</div>	<div>GPU Instance 96GB</div> <div>GPU Instance 96GB</div>	<div>GPU Instance 48GB</div> <div>GPU Instance 48GB</div> <div>GPU Instance 48GB</div> <div>GPU Instance 48GB</div>	<div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div> <div>GPU Instance 24GB</div>

Partitioning mode selected applies to all MI300X GPUs on UBB8 Node

AMD INSTINCT™ MI300X UBB8 GPU VIRTUALIZATION



SINGLE 8GPU 1.5TB VM INSTANCE PER NODE
Large AI Training

EIGHT 1GPU 192GB VM INSTANCES PER NODE
Large AI Inference

KVM HYPERVISOR SUPPORT
Ubuntu Host, Ubuntu Guest OS

SR-IOV VIRTUAL FUNCTIONS

INFINITY FABRIC™ INTERCONNECT SUPPORT

Benefits of AMD Instinct™ Accelerators Partitioning

- GPU partitioning
 - Compute units partitioning
 - GPU memory partitioning

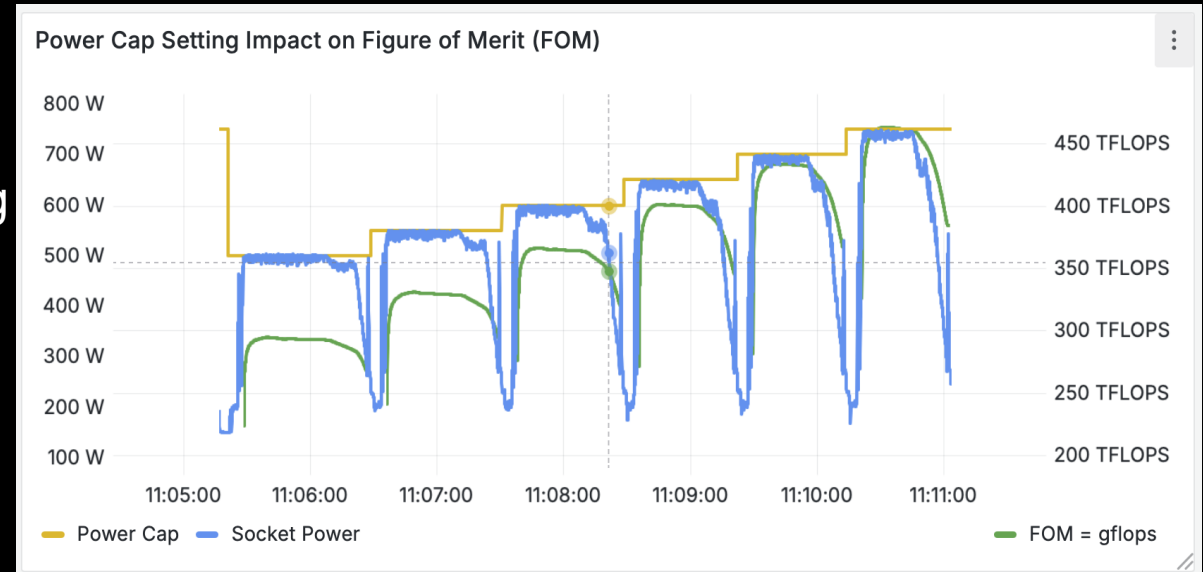


- Benefits
 - Optimizing resource allocation (assigning workloads/tasks to specific partitions)
 - Allowing more efficient resource utilization and reducing the need for whole, more power-hungry GPUs
 - Reducing power consumption

Energy Efficiency on AMD Instinct™ Accelerators

Optimal efficiency combines:

- Benefits of AMD Instinct™ Accelerators partitioning
- Power capping and dynamic adjustments
 - Power budget allocated based on workload demands
 - Dynamic power adjustment per partition
- AI application/benchmark tuning to use the right datatype (FP4 / FP6 datatype instead of FP16 / FP32 for example)
- Application optimization experts



AMD SILO AI

Solving the last mile of customer AI



200+ AI implementations

Helping clients succeed in applying
AI to product development



Open-source base models

European language LLMs trained
on LUMI supercomputer with AMD
Instinct™ accelerators

AMD Sovereign AI Solutions and Energy Benefits

- AMD designs and deploys full-stack AI solutions
 - Hardware
 - Open-source software ; not locking anyone down
 - Models and experts in various AI domains
 - HW/SW co-design experts
- Optimal energy efficiency on AMD Instinct™ Accelerators combines:
 - GPU partitioning
 - Power capping and dynamic adjustments
 - Right datatype usage
 - Application/benchmark tuning/optimization
- AMD is ready for collaboration

Questions?

DISCLAIMERS AND ATTRIBUTIONS

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-18

THIS INFORMATION IS PROVIDED 'AS IS.' AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

© 2025 Advanced Micro Devices, Inc. All rights reserved.

AMD, the AMD Arrow logo, Radeon™, Instinct™, EPYC, Infinity Fabric, ROCm™, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

