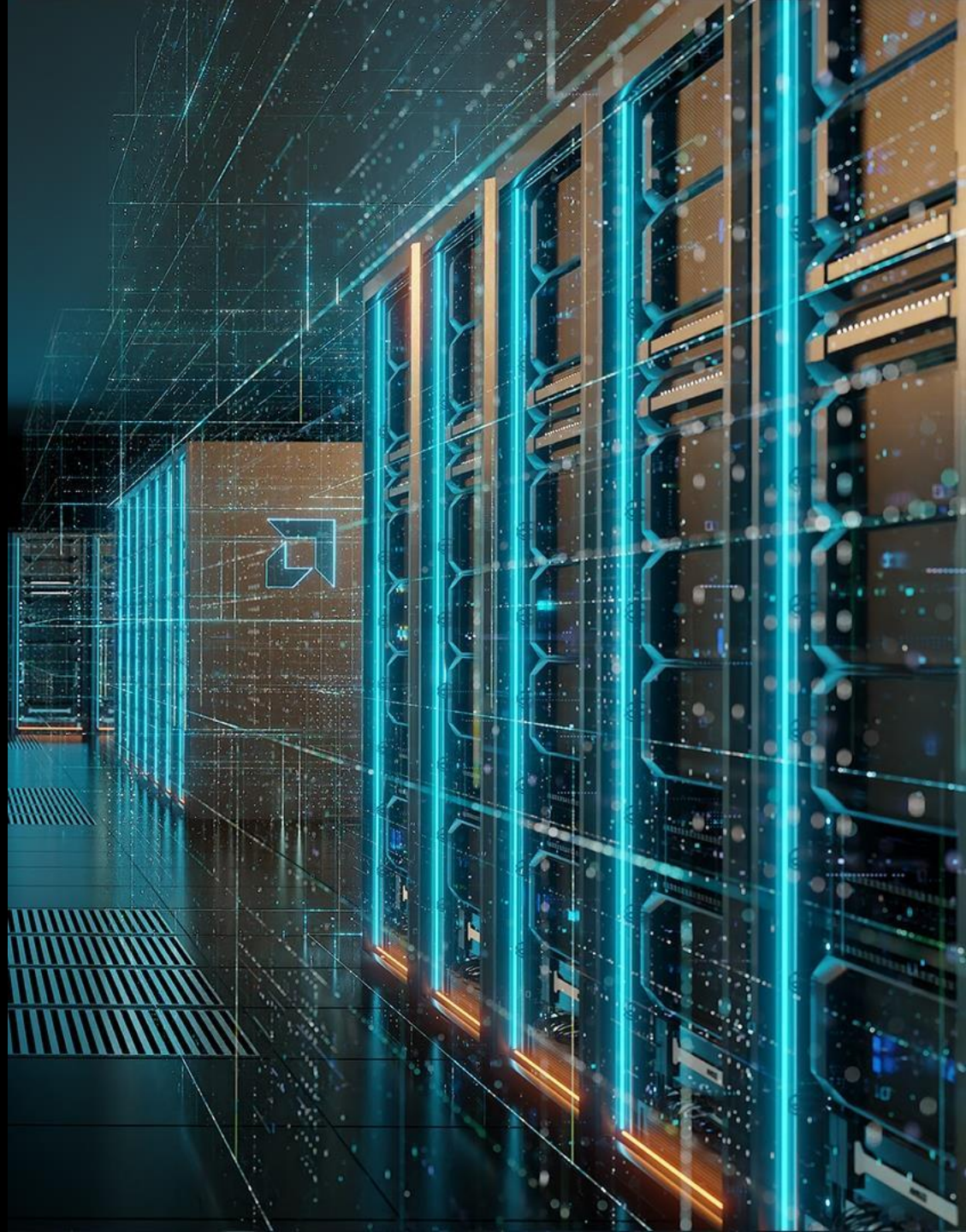
AMD

Forum TERATEC 24

Unlock the future

About the silicon

Jose Noudohouenou, Benjamin Pajot



HIGH-PERFORMANCE COMPUTING: OVERVIEW

What Does HPC require:

- Supercomputers
- Computer clusters (aggregate computing resources)

HPC resources are used to solve advanced computation problems

Where is HPC performed?

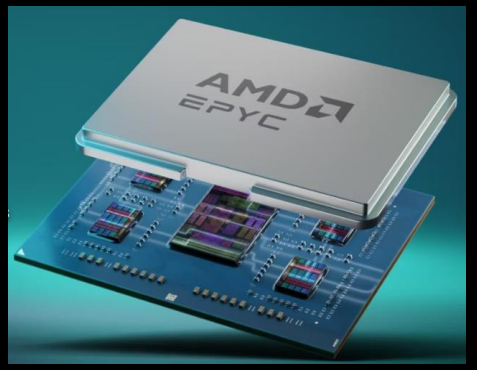
- Traditional HPC: on-premise
- Modern HPC:
  - Cloud computing (offering computing resources for the commercial sector)
  - Hybrid (combine both on-premise and cloud computing)

Key HPC components: compute, storage, and networking

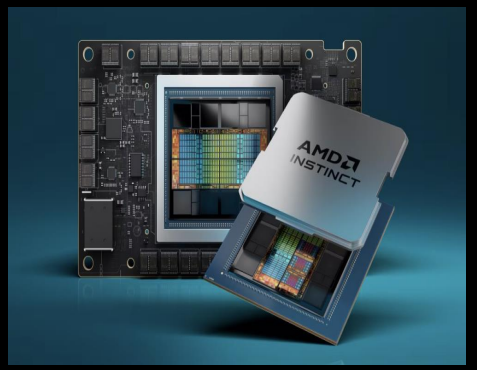
Key actors: Semiconductor companies, Solutions integrators and Cloud service providers, Users

AMD COMPUTE PORTFOLIO

- Server processors
  - AMD EPYC™ processors (CPUs)
    Includes cores with exceptional memory bandwidth and capacity



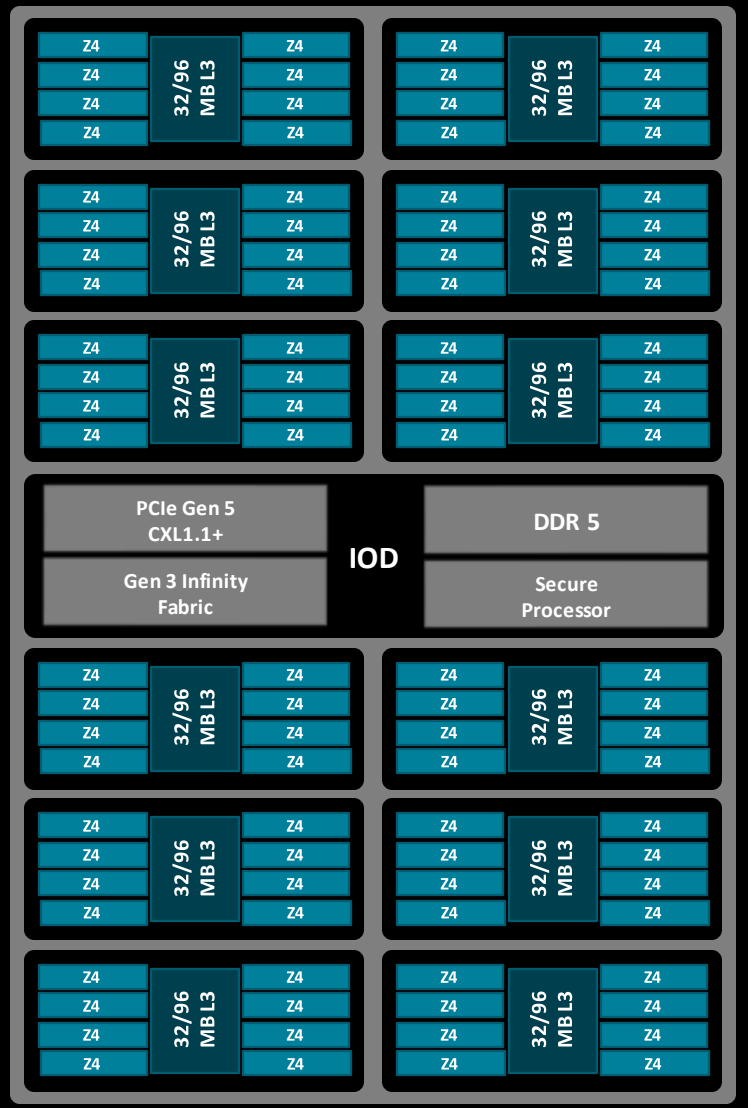
- Accelerators
  - AMD Instinct™ GPUs
    Offer exceptional compute performance, large memory density, high bandwidth memory, and support for specialized data formats.



All with various configurations (depending on needs for both HPC and AI workloads)

AMD EPYC™ 9004 at a Glance

AMD INSTINCT



Up to 96 "Zen 4" cores in 5 nm

ZEN 4

200-360W default TDP

1MB L2 cache per core
32MB of shared L3 cache per CCD

Common socket & platform (SP5)
12 channels DDR5-4800 memory
128 lanes PCIe5

Security Features

Dedicated Security Subsystem with enhancement, Secure Boot, Hardware Root-of-Trust, SME, SEV-ES, SEV-SNP, AES-256-XTS

4th Gen EPYC™ SoC – 97X4

Compute

- AMD "Zen4c" x86 cores (Up to 8 CCDs / Up to 128 cores / 256 threads)
- 1MB L2/Core, 2x 16MB L3 CCX per CCD
- ISA updates: BFLOAT16, VNNI, AVX-512 (256b data path)
- Memory addressability with 57b/52b VA/PA

- Updated IOD and internal AMD Gen3 Infinity Fabric™ architecture with increased die-to-die bandwidth

- TDP range: up to 400W (cTDP)

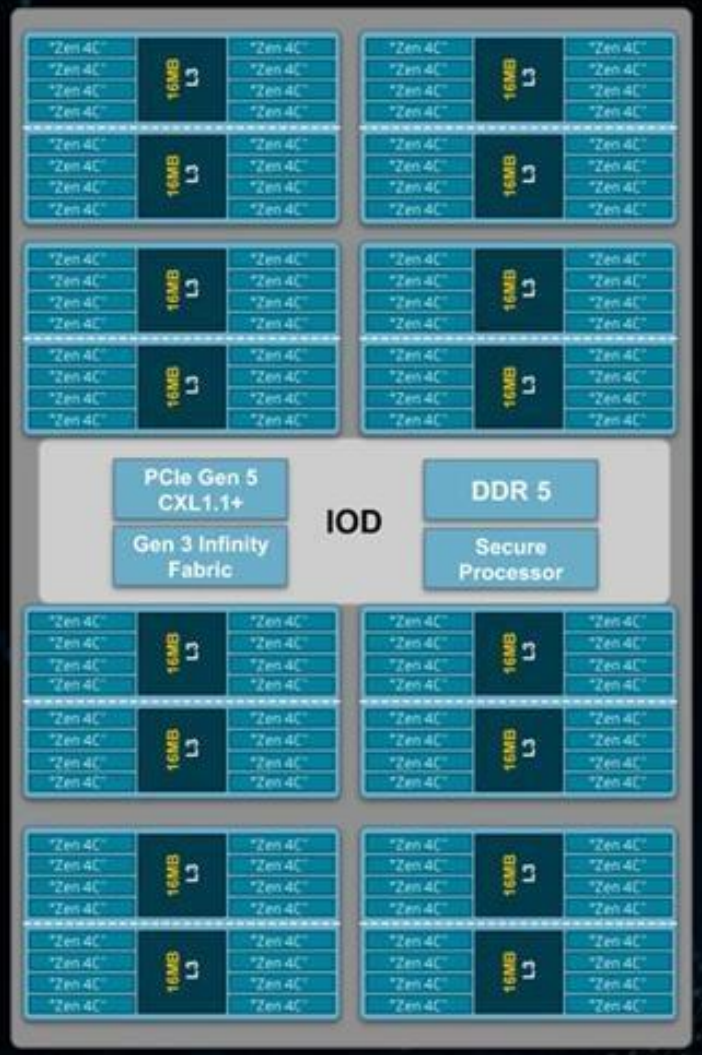
- Updated RAS

Memory

- 12 channel DDR5 with ECC up to 4800 MHz
- Option for 2, 4, 6, 8, 10, 12 channel memory interleaving1

- RDIMM, 3DS RDIMM

- Up to 2 DIMMs/channel capacity with up to 12TB in a 2 socket system (256GB 3DS RDIMMs)1



ORANGE indicates difference from General Purpose

SP5 Platform

- New socket, increased power delivery and VR
- Up to 4 links of Gen3 AMD Infinity Fabric™ with speeds of up to 32Gbps
- Flexible topology options

- Server Controller Hub (USB, UART, SPI, I2C, etc.)

Integrated I/O – No Chipset

Up to 160 IO lanes (2P) of PCIe® Gen5

- Speeds up to 32Gbps, bifurcations supported down to x1
- Up to 12 bonus PCIe® Gen3 lanes in 2P config (8 lanes–1P)
- Up to 32 IO lanes for SATA

- 64 IO Lanes support for CXL1.1+ w/bifurcations supported down to x4

Security Features

Dedicated Security Subsystem with enhancements

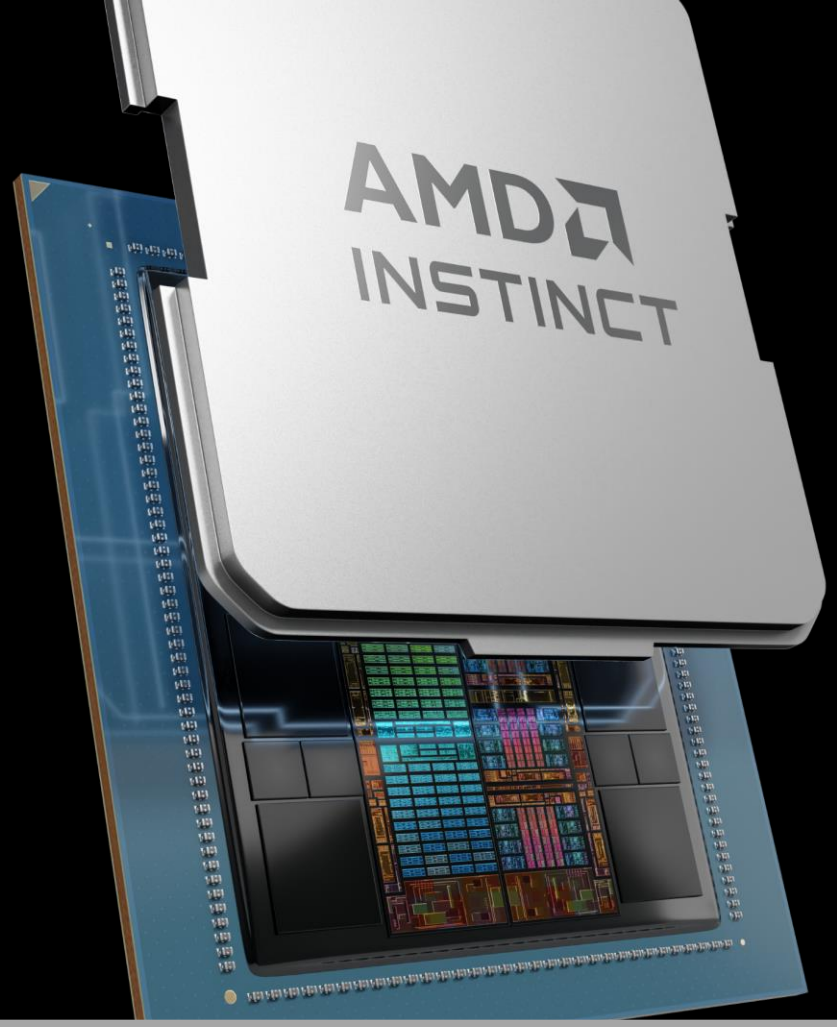
Secure Boot, Hardware Root-of-Trust

SME (Secure Memory Encryption)

SEV-ES (Secure Encrypted Virtualization & Register Encryption)

SEV-SNP (Secure Nested Paging), AES-256-XTS with more encrypted VMs

1 With certain DIMM population rules. GD-183: AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard.



AMD Instinct™ MI300A

World's first data center APU accelerator for AI and HPC

AMD CDNA 3

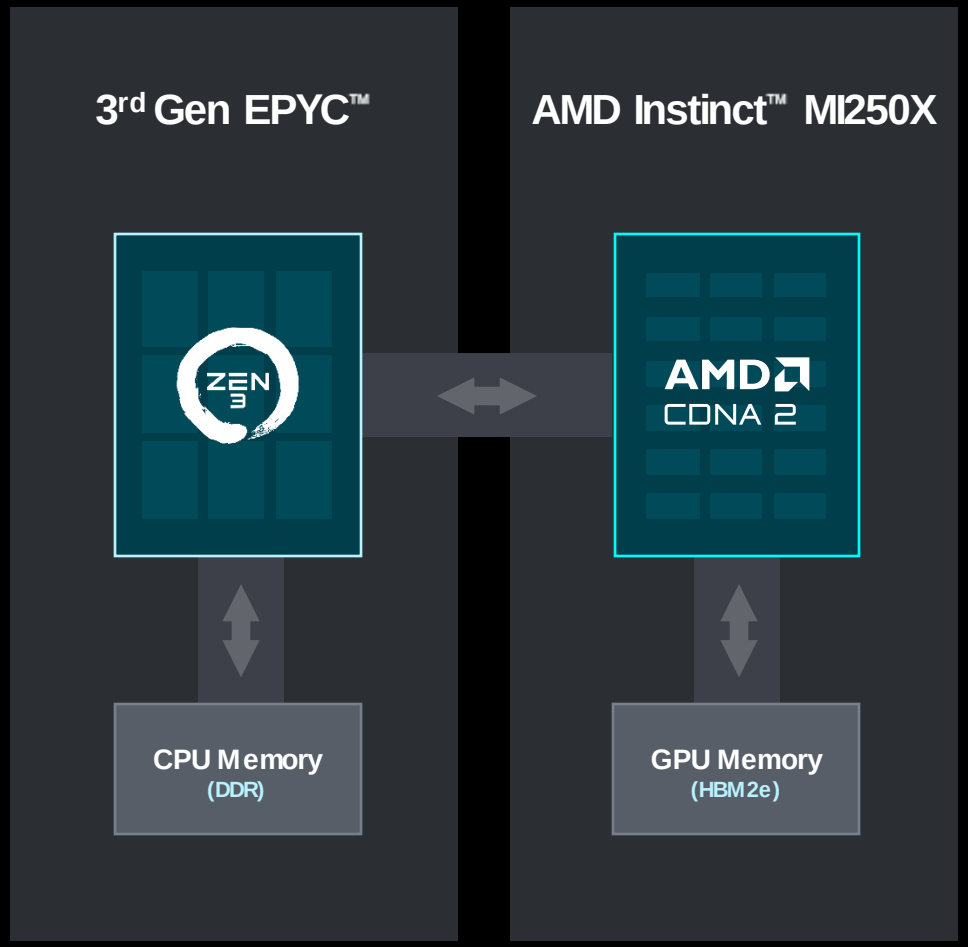
ZEN 4

128 GB
HBM3

5nm and **6**nm
Process Technology

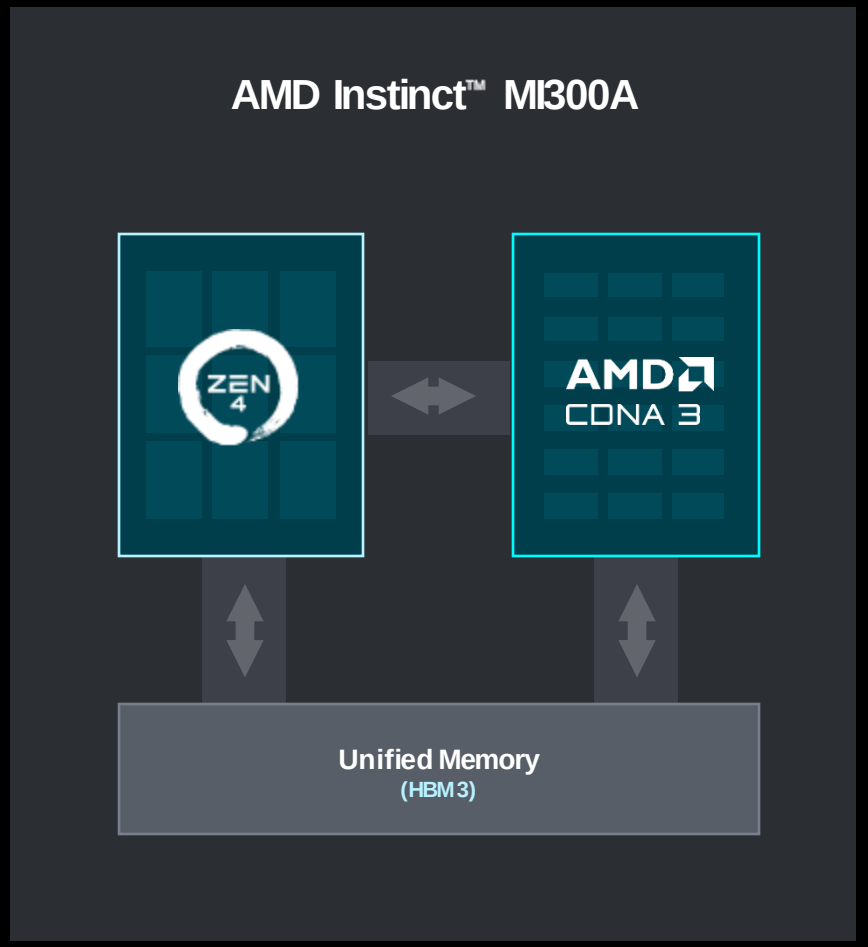
Shared Memory
CPU + GPU

3rd Generation
Infinity Architecture



2021

4th Generation
Infinity Architecture



2023

AMD Instinct™ MI300A Accelerator

AMD INSTINCT

6 XCDs
228 AMD CDNA™ 3 compute units

AMD CDNA 3

3 CCDs
24 “Zen 4” x86 cores CPU

ZEN 4

4 IODs

8 HBM3 stacks

256 MB
AMD Infinity Cache™ technology

3.5D packaging

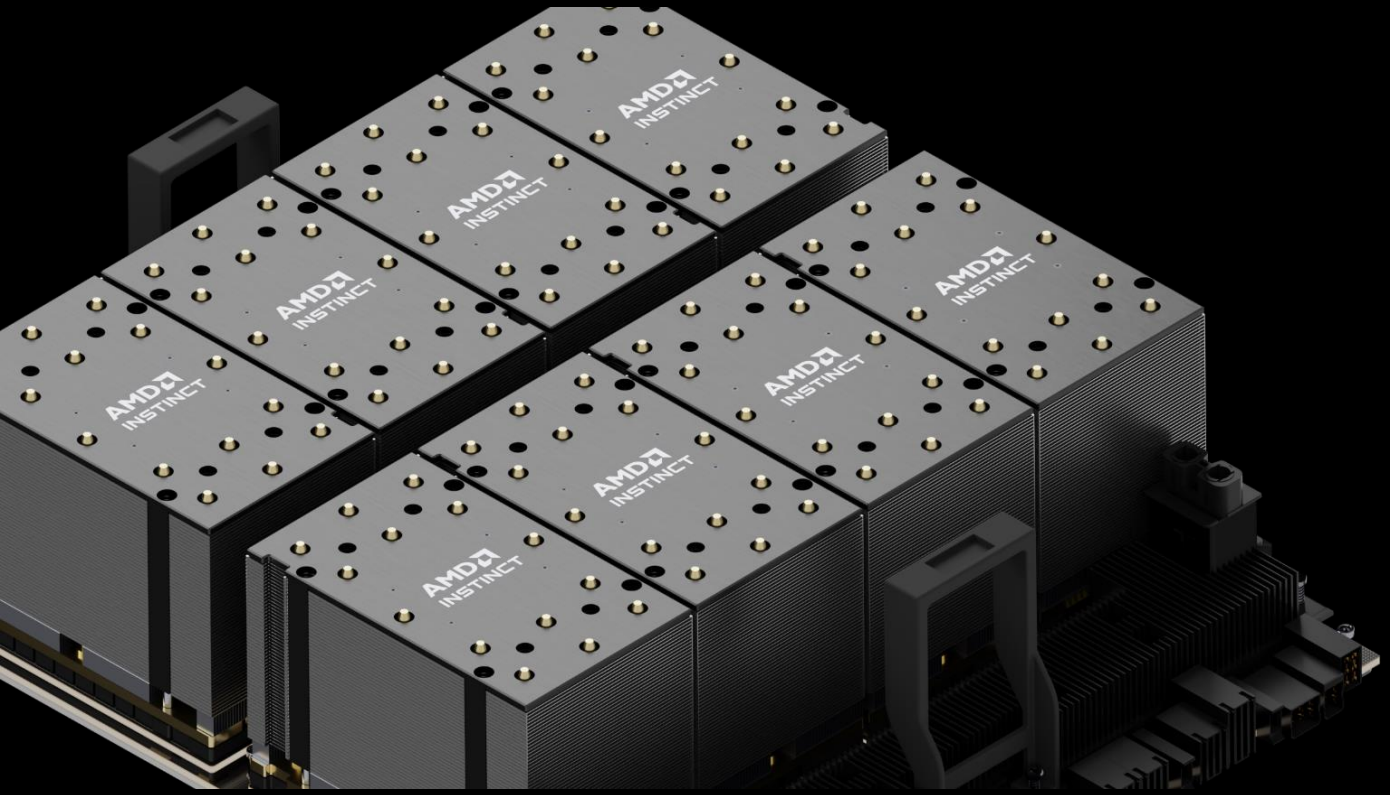
APU Advantage

Unlocking new performance capabilities

Unified Memory	Shared AMD Infinity Cache™ technology	Dynamic Power Sharing	Ease of Programming

AMD Instinct™ MI300X Platform

Industry-leading generative AI platform



8
AMD Instinct™ MI300X

~10.4 PF
BF16/FP16

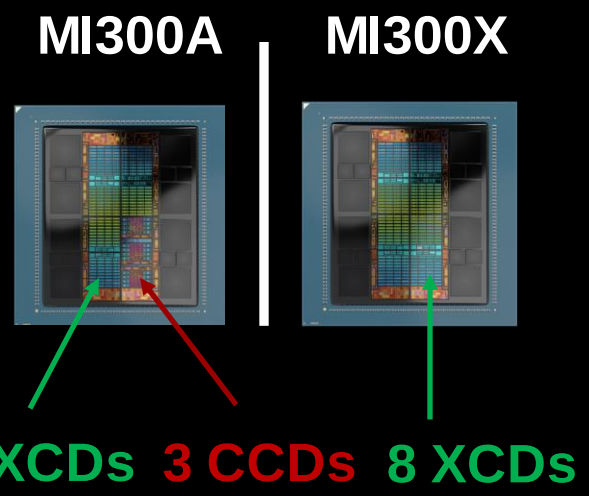
1.5 TB
HBM3

~896 GB/s
Infinity Fabric™ Bandwidth

Industry-Standard Design

MI300APU AND MI300X CHARACTERISTICS

Products	MI300A	MI300X DISCRETE GPU
Max Frequency	2100 MHz	2100 MHz
XCD count per GPU	6	8
CU count per XCD	38	38
L1 Cache Size per CU	32 KB	32 KB
L2 Cache Size per XCD	4 MB	4 MB
AMD “ZEN 4” CPU Chiplets (CCD)	3	N/A
Total “ZEN 4” X86 cores	24	N/A
Total AMD Infinity Cache (LLC)	256 MB	256 MB
Total memory(HBM3) Size	128 GB	192 GB
Stream processors	14 592	19 456
Memory Bandwidth (peak)	up to 5.3 TB/sec	up to 5.3 TB/sec
Thermal	Passive & Liquid	Passive & Liquid
Max Power	550W or 760W	750W



XCD: Accelerated Complex Die
CCD: CPU Complex Die

THEORETICAL PEAK COMPUTE PERFORMANCE COMPARISON MI300APU AND MI300X DISCRETE GPU

Computation	MI300A GPU (Peak TFLOP/s)	MI300X GPU (Peak TFLOP/s)
FP64 VECTOR	61.3	81.7
FP32 VECTOR	122.6	163.4
FP64 MATRIX	122.6	163.4
FP32 MATRIX	122.6	163.4
TF32 MATRIX | TF32 (SPARSITY)	490.3 | 980.6	653.7 | 1 307.4
FP16 | FP16 (SPARSITY)	980.6 | 1 961.2	1 307.4 | 2 614.9
BF16 | BF16 (SPARSITY)	980.6 | 1 961.2	1 307.4 | 2 614.9
FP8 | FP8 (SPARSITY)	1 961.2 | 3 922.3	2 614.9 | 5229.8
INT8 | INT8 (SPARSITY)	1 961.2 TOPs | 3 922.3 TOPs	2 614.9 TOPs | 5 229.8 TOPs

TFLOPS: Trillions Floating Point Operations per Second
TOPS: Trillions Operations per Second

HPC KEY ACTORS

Key end-to-end actors:

1. Semiconductor companies
   Design and manufacture semiconductor devices like microprocessors used in supercomputers

2. Solutions Integrators and Cloud Service Providers
   Either integrate or develop semiconductor companies' products and other components to offer computing resources on the market

3. Users (including scientists)

Each of these actors has challenges and opportunities especially when it is about computing resources or moving to Cloud computing

EXAMPLE OF CLOUD SERVICES
AMD ACCELERATOR CLOUD (AAC)

- Private AMD cloud environment that enables a remote access to AMD hardware and software technologies
  - Hardware already setup
  - Software already installed by AMD engineers
    - Easy to deploy containers available for all sorts of workloads and environments
    - ROCm already preconfigured

- Purposes:
  - Try it before you buy it
  - Live, Interactive trainings & Demos

- Targets:
  - Existing customers
  - Prospective buyers
  - Learners and experts (hackathon and training events)

EXAMPLE OF CLOUD SERVICES: AAC (2)

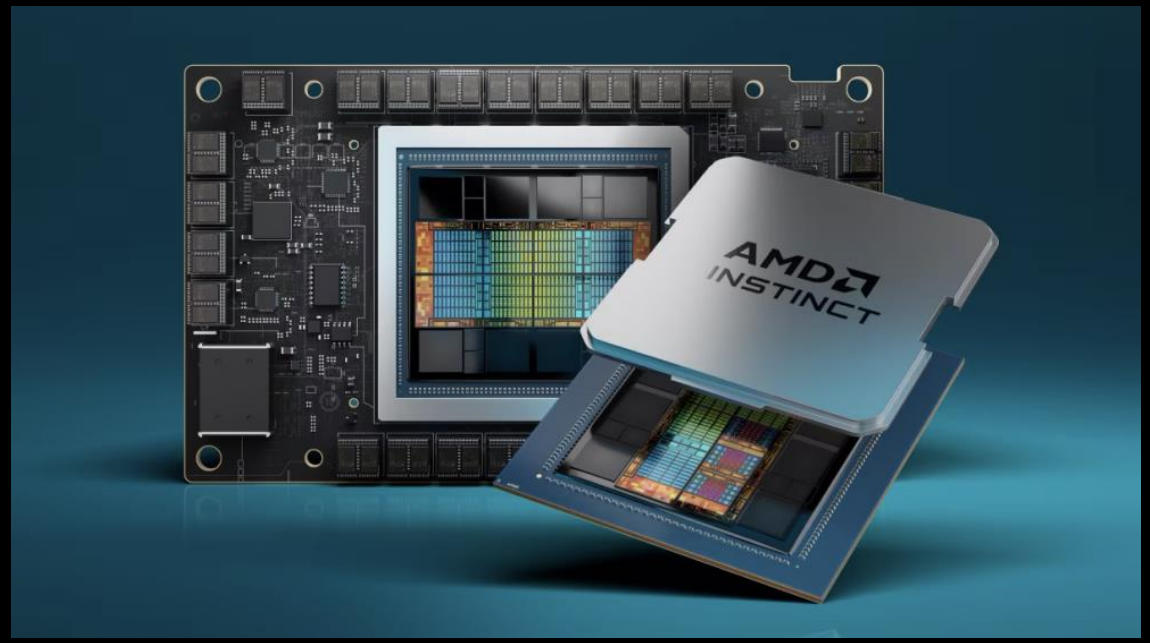
- Limited time offers
  - Dedicated machines (private)
  - Shared machines
- Product Categories
  - Compute nodes
  - Containers (Plexus)
  - Kubernetes (AI workloads)
- Accepted workload types: HPC and AI

EXAMPLE OF CLOUD SERVICES: AAC (3)

- Don’t
  - No subscription is needed
- Do
  - An account on AAC node is needed
  - Share your public ssh key

AAC COMPUTE NODES (1)

- Compute node types:
  - AMD EPYC™ CPUs
  - AMD Instinct™ GPUs
  - Etc..



AMD compute nodes include existing as well as early AMD machines (or machines that are barely out - you won't get some of these machines from general cloud service providers)

AAC COMPUTE NODES (2)

Current compute node instances

Instances	Types	Size (Based on priorities)
CPUs	AMD EPYC 32-Core Processor AMD EPYC 48-Core Processor AMD EPYC 64-Core Processor AMD Instinct MI300A Accelerator Etc…	Sampling of each
AMD Instinct GPU	MI300X MI300A (APU) MI250 MI210 MI100	Sampling of each

OS:

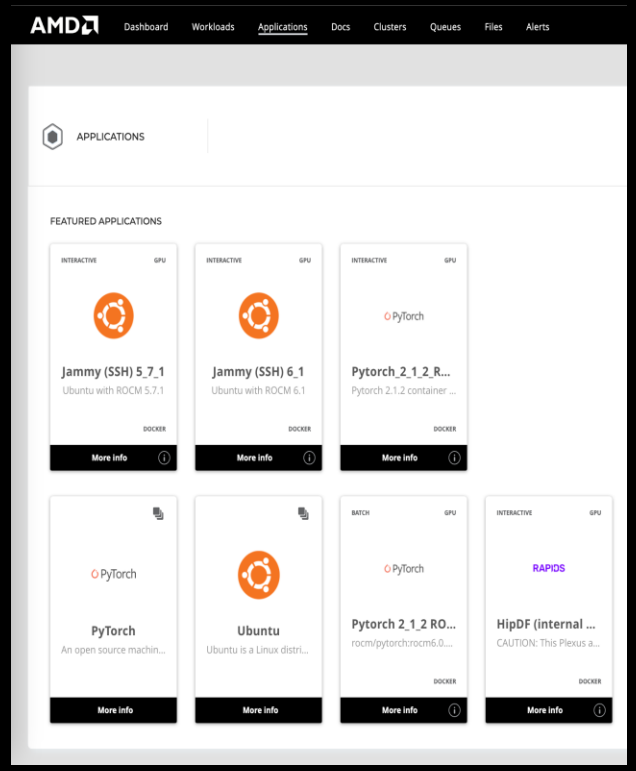
- Ubuntu
- RedHat
- SLES

- Software: AMD ROCm™, AMD profiling tools, Cray Software, Pytorch, TensorFlow, Kubernetes, etc…

Variety of node configurations with a range of AMD Instinct accelerator options.
For the moment, MI300X only supports Ubuntu.

AAC CONTAINERS

- Containers: Setup for various purposes
  - Demos
  - Trainings (On-Demand & Private trainings; Virtual trainings)
  - Hackathons
  - Etc…



TRAININGS

- On-Demand & Private Trainings
- Virtual trainings for groups -- EuroHPC sites periodically sponsor events; others can request
- Hackathons -- more hands-on work with your applications; on-site and virtual for groups with AMD hardware
- Office Hours -- offered at some HPC sites; Get support for issues you are encountering

Note: Training events are limited by available resources and sites for events

RESOURCES

- AAC Guide/Docs is at: https://github.com/amddcgpuce/AMDAcceleratorCloudGuides/tree/main/AACPlanoSlurmCluster

- AMD engineers available to support/help you for:
  - Onboarding
  - Learning
  - Application tuning & optimization

Experience the power of the AMD solution! – Play with your application in a real environment.

CONTACTS

Do you plan to access AMD Accelerator Cloud?
Please, talk to any AMD employee and he/she will put you in contact with an AMD sale representative or BD

For hackathon and training events, contact your AMD Training team
Bob.robey@amd.com
Jose.noudohouenou@amd.com

DISCLAIMER

AMD, the AMD Arrow logo, EPYC™, Instinct™ 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.

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

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.