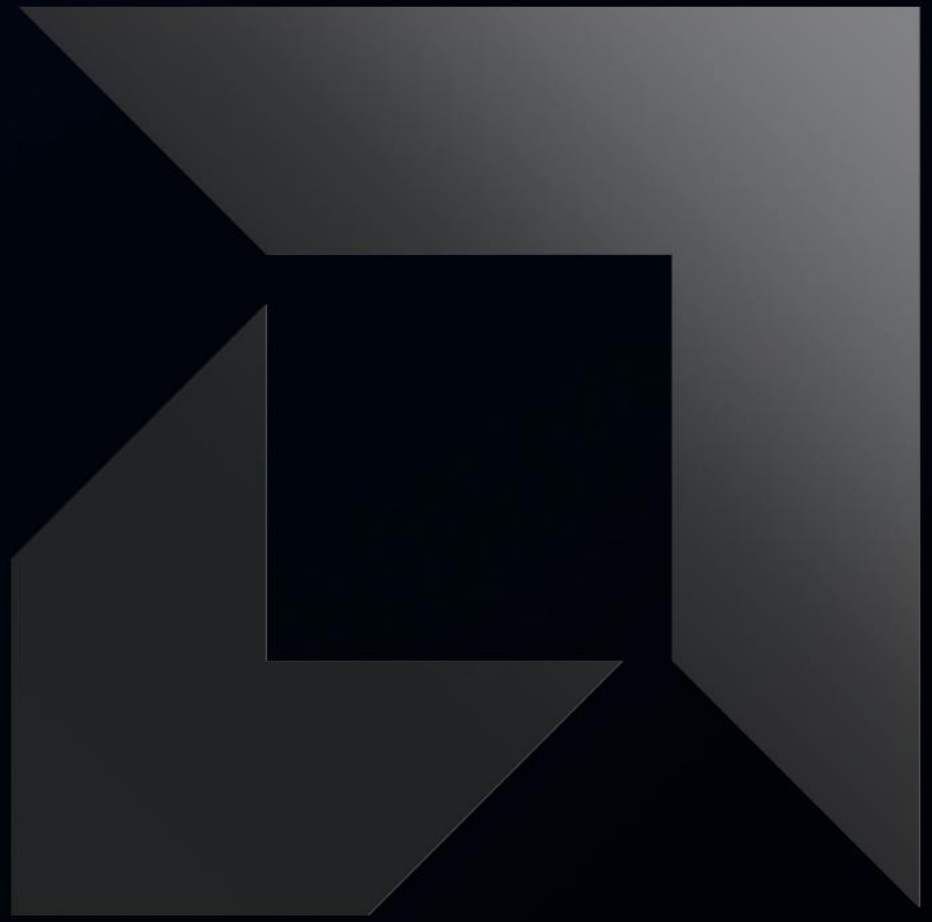


Architecture overview – APU vs Discrete GPU

Bob Robey

UCL

DiRAC TRAINING ACADEMY

AMD together we advance_

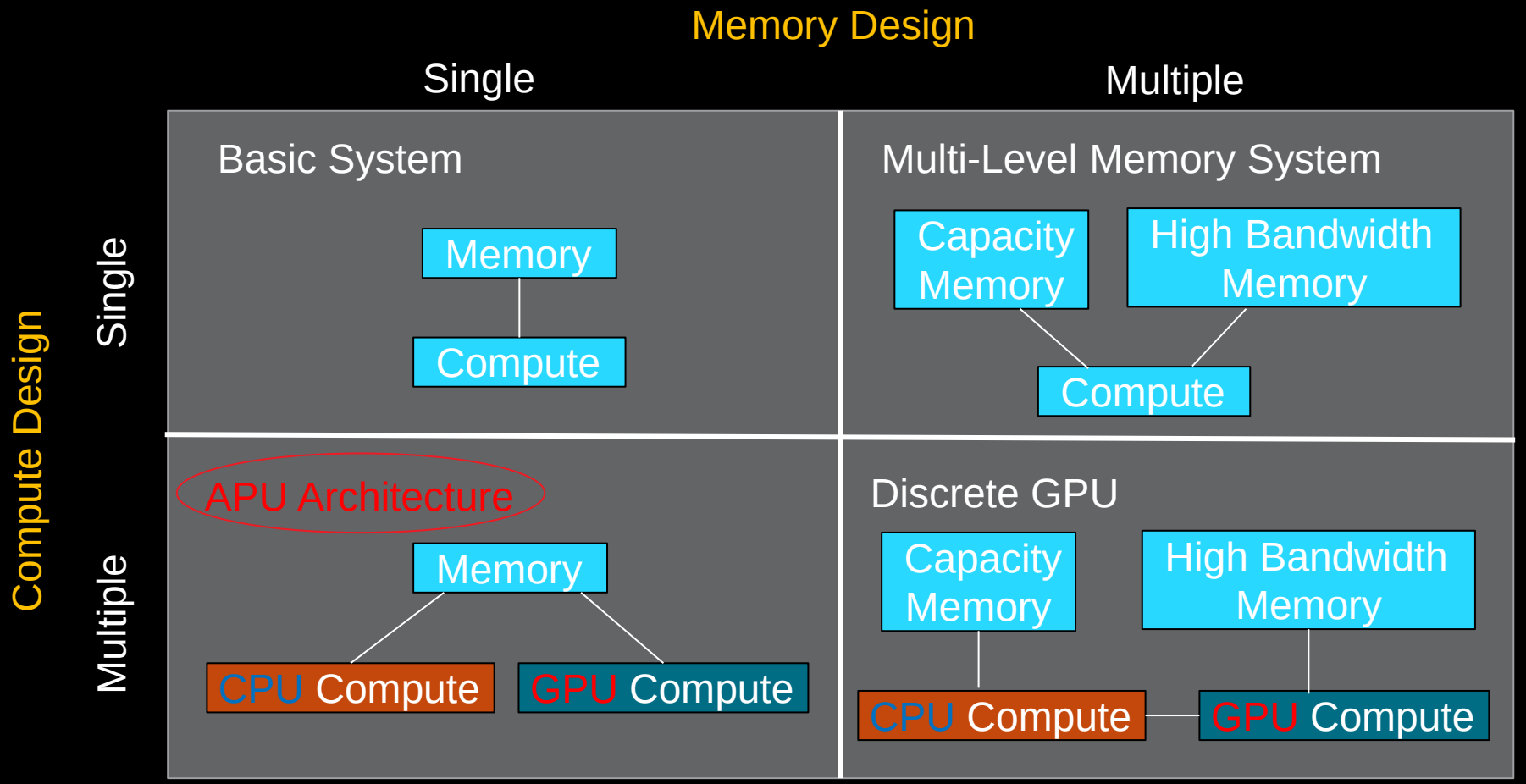
What is covered in this presentation

- How to categorize different CPU and GPU architecture configurations
- How to pick which architecture works best for your applications
- Evaluate which architecture is most cost effective for your application workload
- Understand the architectural underpinnings of the AMD Instinct GPUs (at a high level)

Terminology

- APU – Accelerated Processing Unit, originally an AMD marketing term for the CPU with integrated graphics formed from the merge of the AMD CPU and the ATI GPU onto one chip.
- Integrated graphics – refers to CPUs that have built-in GPU support circuitry.
- Discrete GPU – a GPU that is separate from the CPU and generally installed as a PCIe card.
- High Bandwidth Memory (HBM) - a three-dimensional stacked memory architecture with "vias" or "passthroughs" through the silicon layers that provides dramatically higher memory bandwidth through parallel retrievals from each layer.
- Double Data Rate (DDR) memory – the most common memory for computer systems today
- Chiplets – production technique that joins together smaller chips into one cohesive unit. Large chip size can often lead to low yield of "good" chips. Chiplets can improve yield of the combined chip and also provide quicker adaptability of the chip product.
- Heterogeneous Memory Management (HMM) - management of memory between two diverse compute devices.
- Virtual Memory – the memory page list that gets mapped to physical memory upon demand. Originally this allowed the computer to look like it had more memory by swapping pages and applications into physical devices and support multi-user and multiple memory intensive applications to run at the same time.
- Physical Memory – actual memory on chips.

Taxonomy of compute architectures

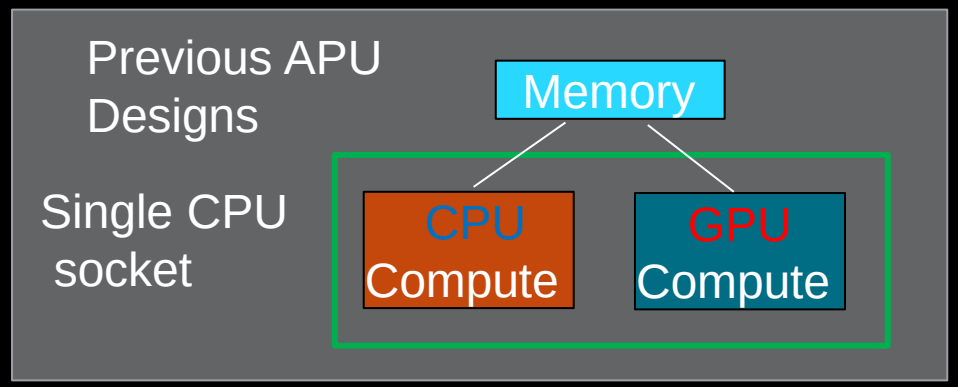


Taxonomy categorizes architecture by dominance of hardware components

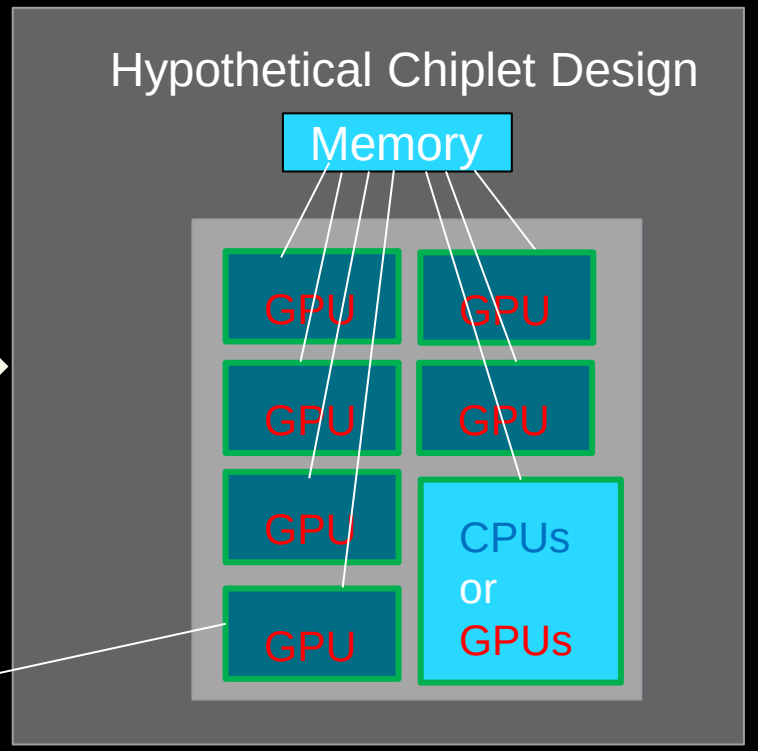
- Memory Dominant – architecture revolves around a single memory space
- Compute Dominant – architecture centered around a single compute resource
- APU is primarily characterized by compute units being able to address all memory

Breakthrough in compute capability -- Conceptual

- Integrated GPUs have traditionally been limited by how much GPU compute capability can be included
  - ❖ Silicon Chip only has so much space
  - ❖ Chiplets allow us to expand that space
- Let's try adding more capability into an APU



More compute capability for design with chiplets



AM5 socket 40x40mm – limited silicon space

Lots more silicon space to work with

UCL

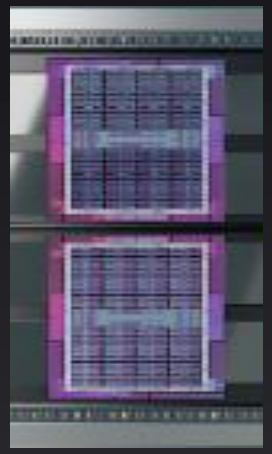
DiRAC TRAINING ACADEMY

Architecture

AMD together we advance_

Bringing it to AMD Instinct™ Accelerator Products

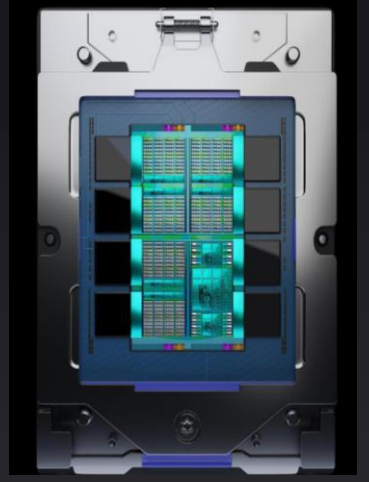
MI200 Series



Extreme Compute Architecture with leading memory capacity and bandwidth

- Technology in first Exascale systems
- High compute to power ratio
- Tight integration with memory
- Infinity Fabric™ for data transfers

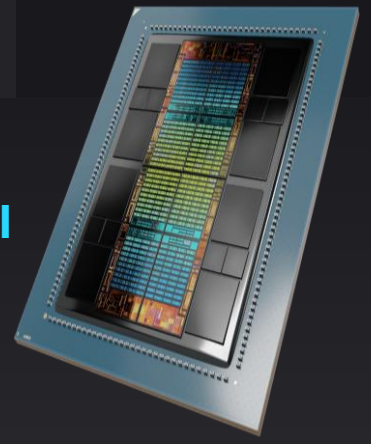
MI300A



First True APU Architecture for HPC and AI

- Memory Bandwidth Workloads
- Hybrid CPU + GPU Capability
- GPUs can drive full bandwidth

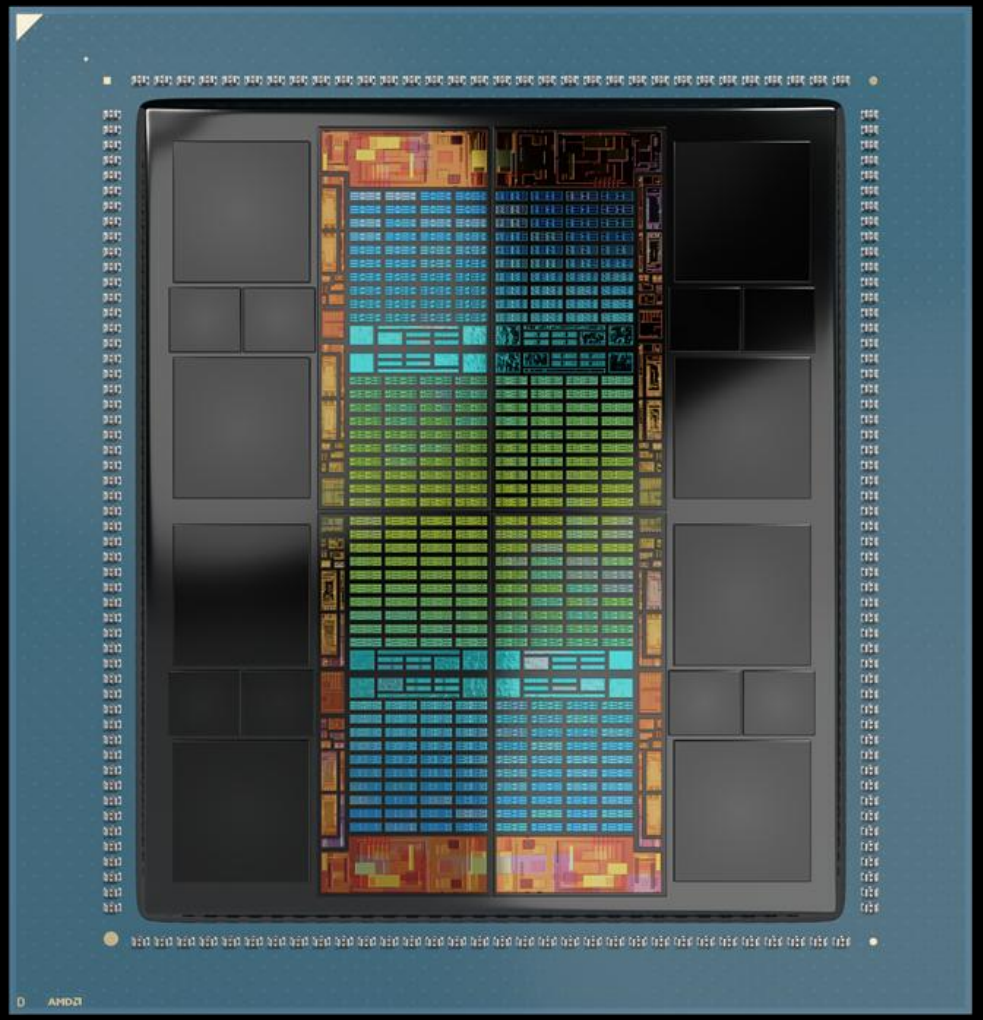
MI300X



Leadership generative AI accelerator

- Extreme Compute Workloads
- Suitable for typical AI work
- Other work entirely on GPU

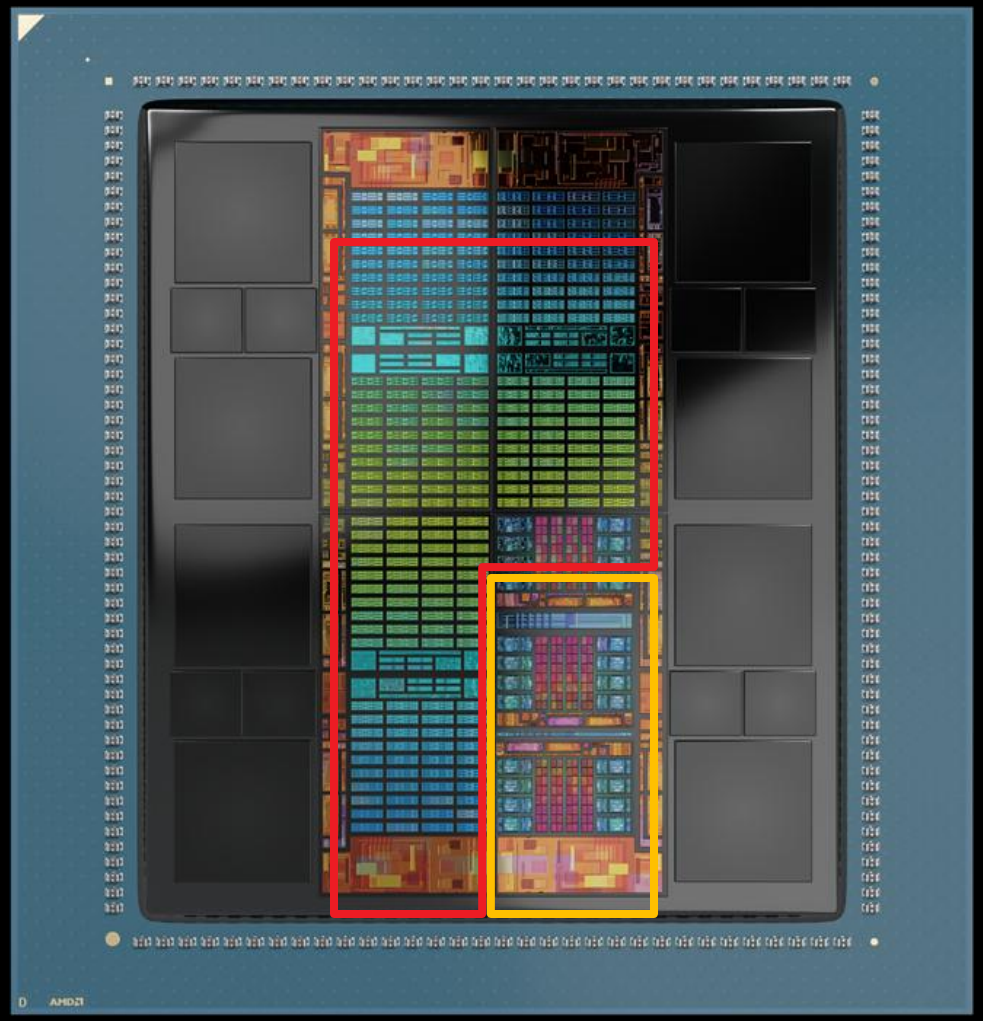
AMD CDNA™ 3 Architecture – AMD Instinct™ MI300X



OAM: OCP Accelerator Module
OCP: Open Compute Platforms

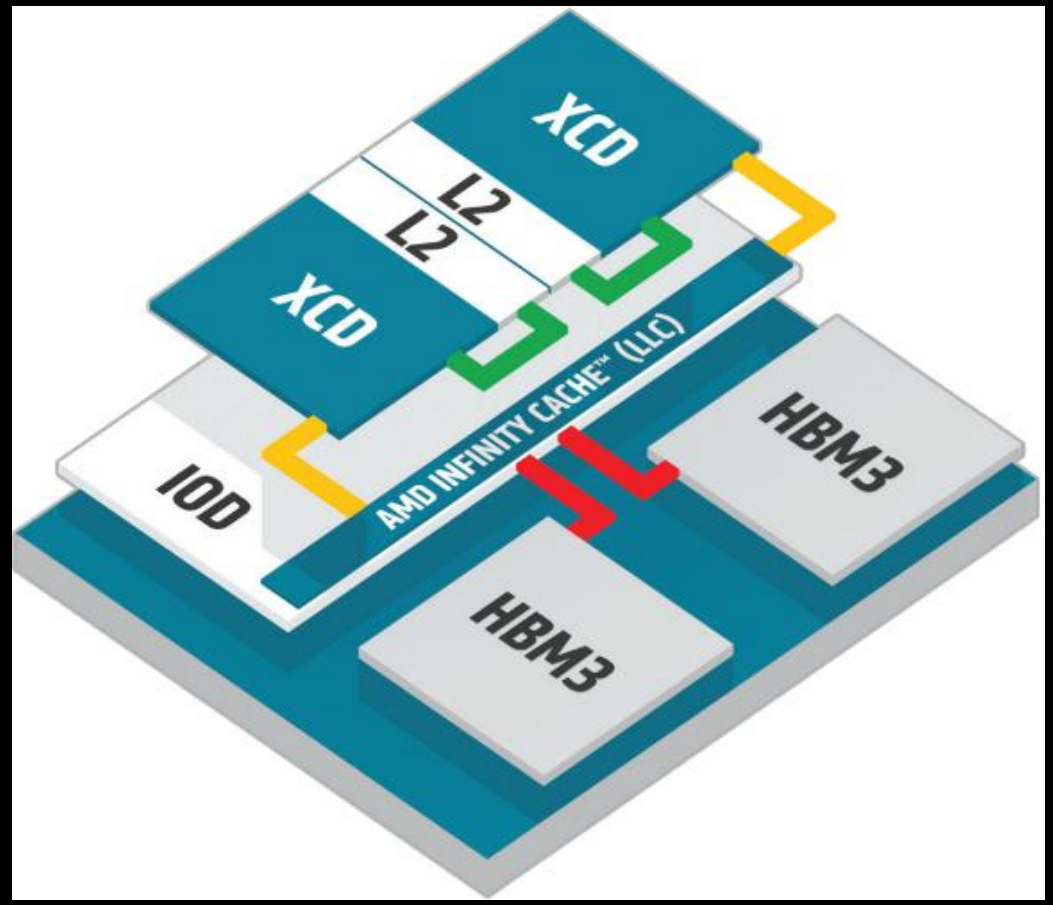
- Discrete GPU
- 304 AMD CDNA™ 3 compute units
- 192 GB HBM3
- Discrete GPU architecture, OAM
  - AMD EPYC™ Processor as the host system
  - Use host system for memory capacity beyond GPU HBM capacity
  - Connects
    - to host via PCIe® Gen 5
    - to other GPUs via AMD Infinity Fabric™ Links
- Supports unified shared memory via page migration (PCIe)

AMD CDNA™ 3 Architecture – AMD Instinct™ MI300A

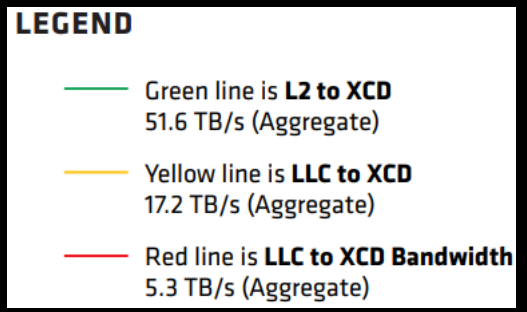


- APU with unified shared memory
- 3D chiplet packaging
- 24 cores, “Zen 4” architecture (CPU), 3 Compute Core Dies (CCDs)
- 228 AMD CDNA™ 3 compute units (GPU), 6 Accelerated Compute Dies (XCDs)
- 128 GB HBM3, unified memory architecture
  - no discrete CPU and GPU memory
  - CPU and GPU access same physical memory

AMD CDNA™ 3 ARCHITECTURE MEMORY ARCHITECTURE DIAGRAM

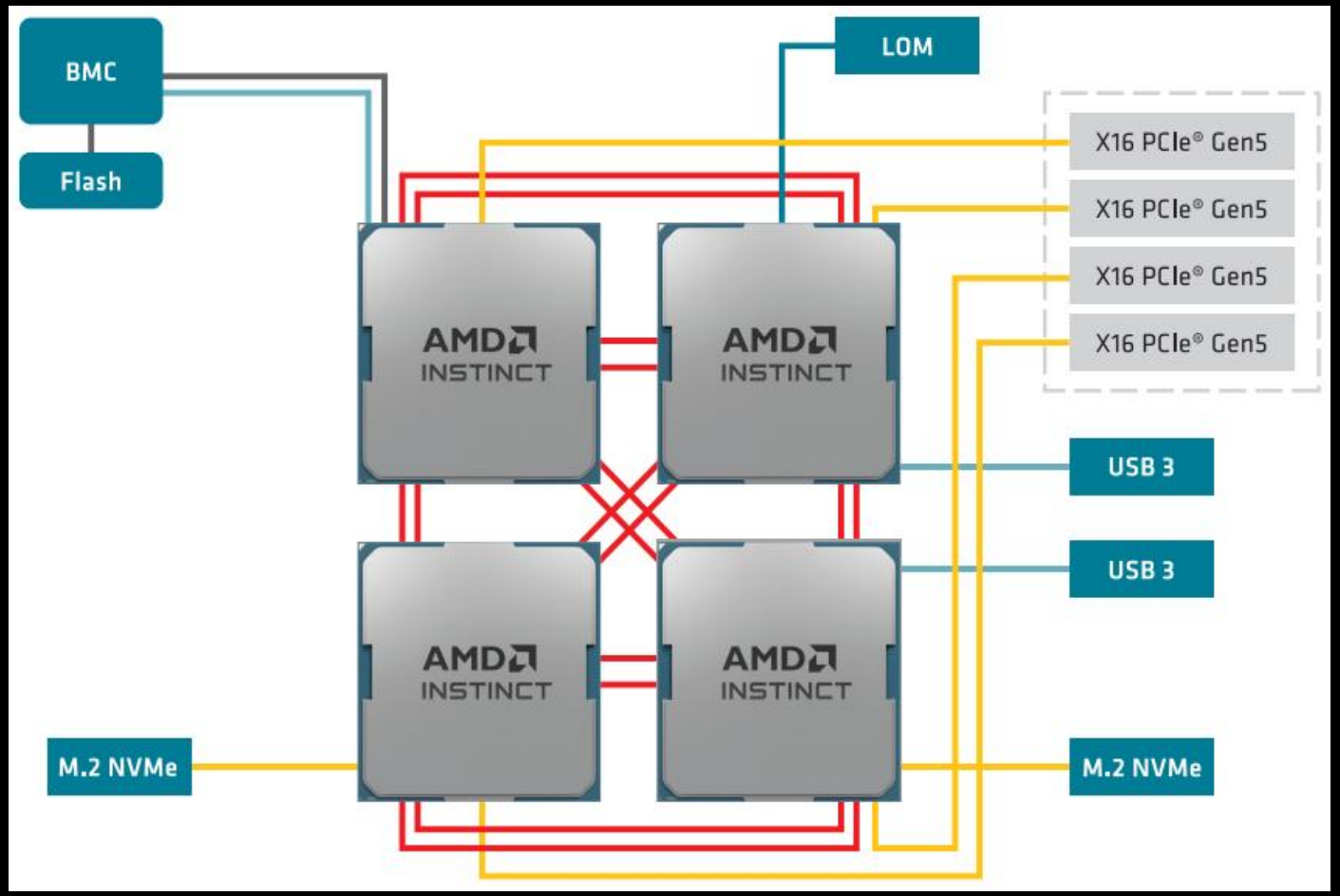


XCD: Accelerated Complex Die
IOD: I/O dies
HBM: High-Bandwidth Memory
LLC: Last Level Cache



CDNA3 whitepaper: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

AMD INSTINCT™ MI300A: 4-APU NODE EXAMPLE



AMD Instinct™ MI300A APU platform design with 4 APUs connected with AMD Infinity Fabric™ links

APU-APU 128 GB/s bi-direction

MI250X AND MI300APU CHARACTERISTICS

Product	MI250x
Max Frequency	1700 MHz
GCD count per GPU	2
CU count per GCD	110
L1 Cache Size per CU	16 KB
L2 Cache Size per GCD	8 MB
Total Memory(HBM2e) Size	128 GB
Stream processors	14 080
Memory Bandwidth (peak)	up to 3.2 TB/sec
Thermal	Passive & Liquid
Max Power	500W & 560W TDP

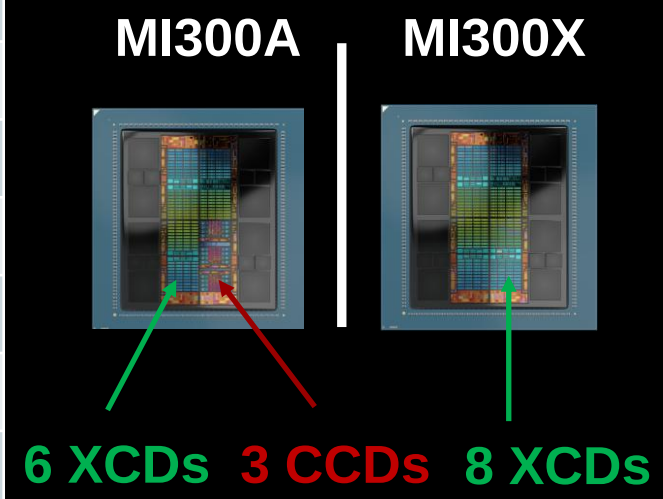
GCD: Graphics Complex Die
TDP: Thermal Design Power

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

XCD: Accelerated Complex Die
CCD: CPU Complex Die

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 COMPARISON: MI300A APU 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

Benefits of APU architecture

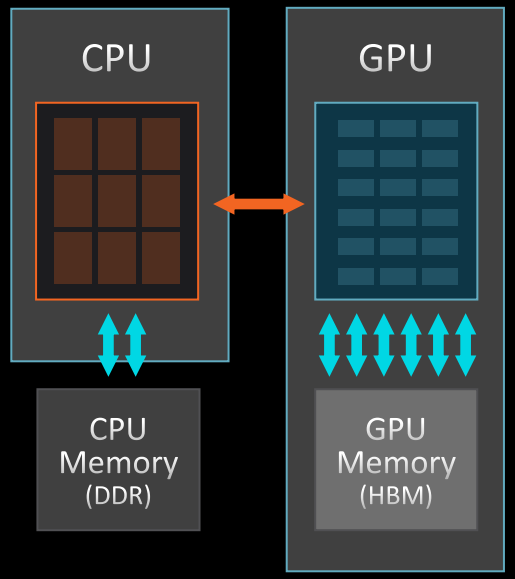
- **Memory is the critical hardware component for today's architectures**

- **Cost**
  - Memory is historically 50% of the cost of the system
  - Duplicating memory adds substantially to cost
  - Another viewpoint is that it halves the memory available to apps

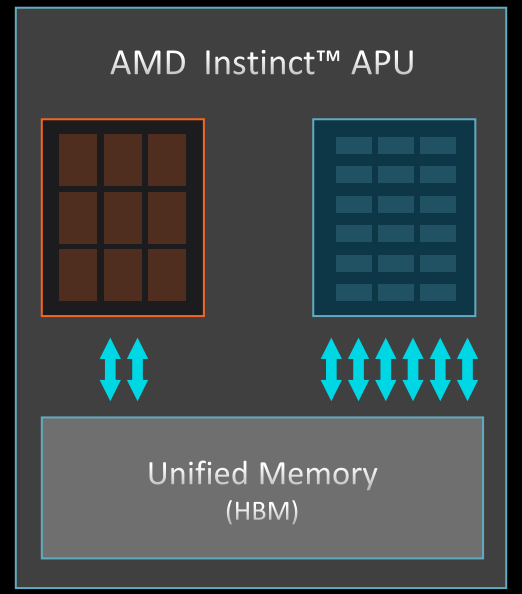
- **Power**
  - Memory consumes a large portion of power demand. Duplicating memory spaces increases the power consumption.
  - Movement between memory spaces greatly increases power draw

AMD MI300A UNIFIED MEMORY APU ARCHITECTURE BENEFITS

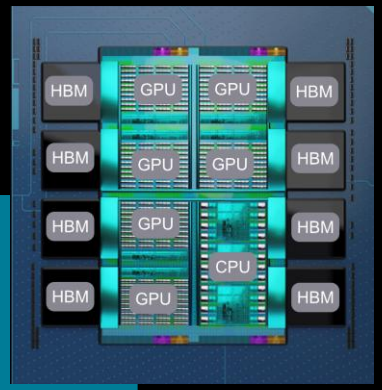
AMD CDNA™ 2 Coherent Memory Architecture → AMD CDNA™ 3 Unified Memory APU Architecture



- **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**



APU Advantage: unified virtual and physical memory



Unified *Virtual* Memory Addressing is supported on systems with discrete devices (and discrete memories).
Heterogeneous Memory Management supports Unified Virtual Memory Addressing

Docs » Linux Memory Management Documentation » Heterogeneous Memory Management (HMM) View page source

Heterogeneous Memory Management (HMM)

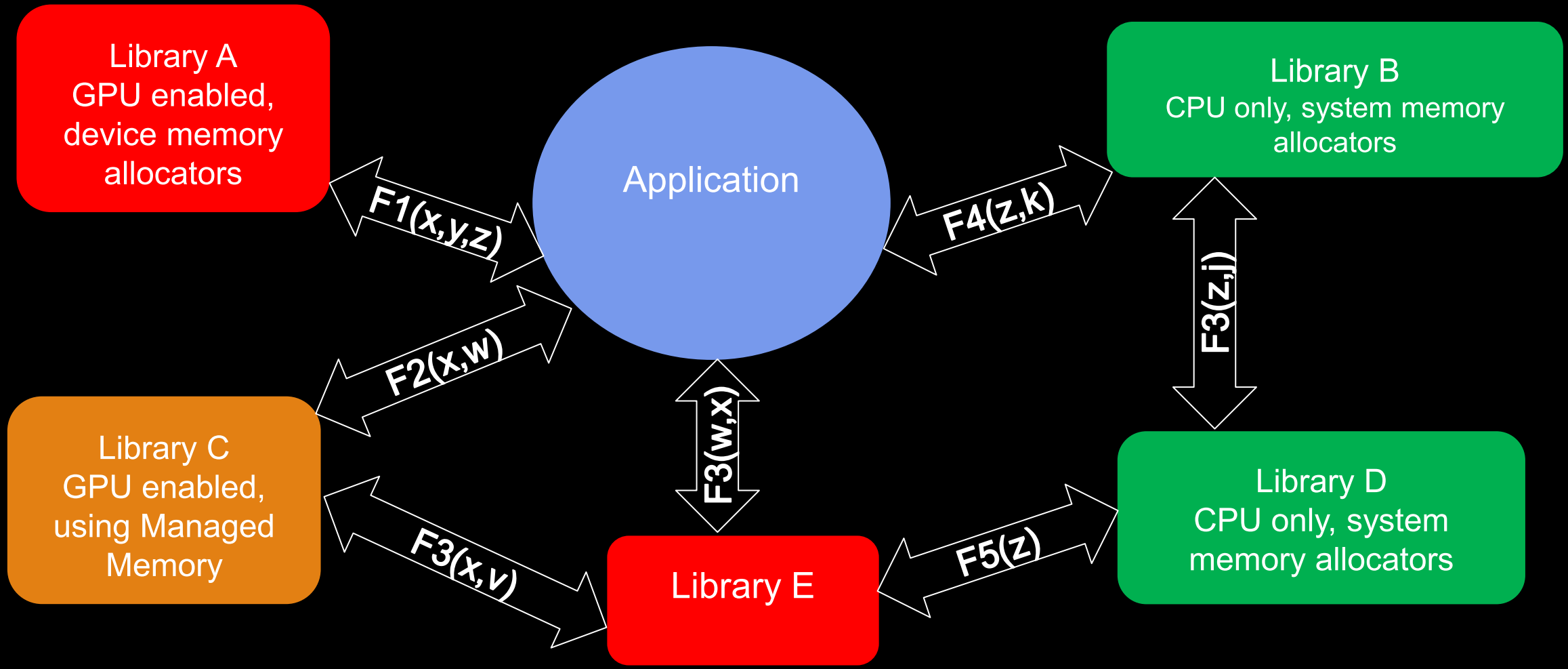
Provide infrastructure and helpers to integrate non-conventional memory (device memory like GPU on board memory) into regular kernel path, with the cornerstone of this being specialized struct page for such memory (see sections 5 to 7 of this document).

HMM also provides optional helpers for SVM (Share Virtual Memory), i.e., allowing a device to transparently access program address coherently with the CPU meaning that any valid pointer on the CPU is also a valid pointer for the device. This is becoming mandatory to simplify the use of advanced heterogeneous computing where GPU, DSP, or FPGA are used to perform various computations on behalf of a process.

https://www.kernel.org/doc/html/v5.0/vm/hmm.html

MI300A offers Unified *Virtual* and Unified *Physical* Memory, which preserves the programmability aspect and enhances performance via eliminating an overhead of managing data in different physical memory spaces (i.e. page migrations, or explicit pinning of pages to a particular device memory, etc.)

Unified memory simplifies data interoperability between an application and libraries



Disclaimer

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.

Third-party content is licensed to you directly by the third party that owns the content and is not licensed to you by AMD. ALL LINKED THIRD-PARTY CONTENT IS PROVIDED “AS IS” WITHOUT A WARRANTY OF ANY KIND. USE OF SUCH THIRD-PARTY CONTENT IS DONE AT YOUR SOLE DISCRETION AND UNDER NO CIRCUMSTANCES WILL AMD BE LIABLE TO YOU FOR ANY THIRD-PARTY CONTENT. YOU ASSUME ALL RISK AND ARE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY ARISE FROM YOUR USE OF THIRD-PARTY CONTENT.