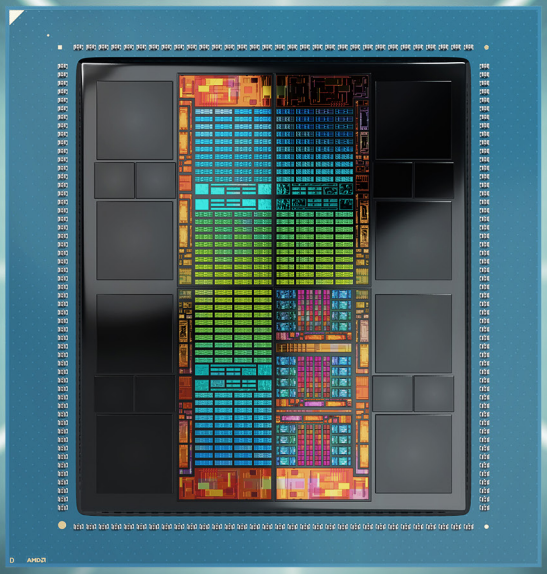


AMD CDNA™ 3 ARCHITECTURE

The All-New AMD GPU Architecture for the Modern Era of HPC and AI

AMD ROCm

AMD INSTINCT

AMD together we advance_

AMD INSTINCT

INTRODUCTION

Over the past three decades, GPUs have revolutionized the entire computing ecosystem, from smartphones and PCs to data centers. Some of the most exciting applications today, enabled by machine learning, would not be possible without GPU acceleration. Over this time, the humble graphics card has transformed from a fixed-function offload engine to a general-purpose processor that accelerates programming languages such as Python, C++, and Fortran - and its evolution continues.

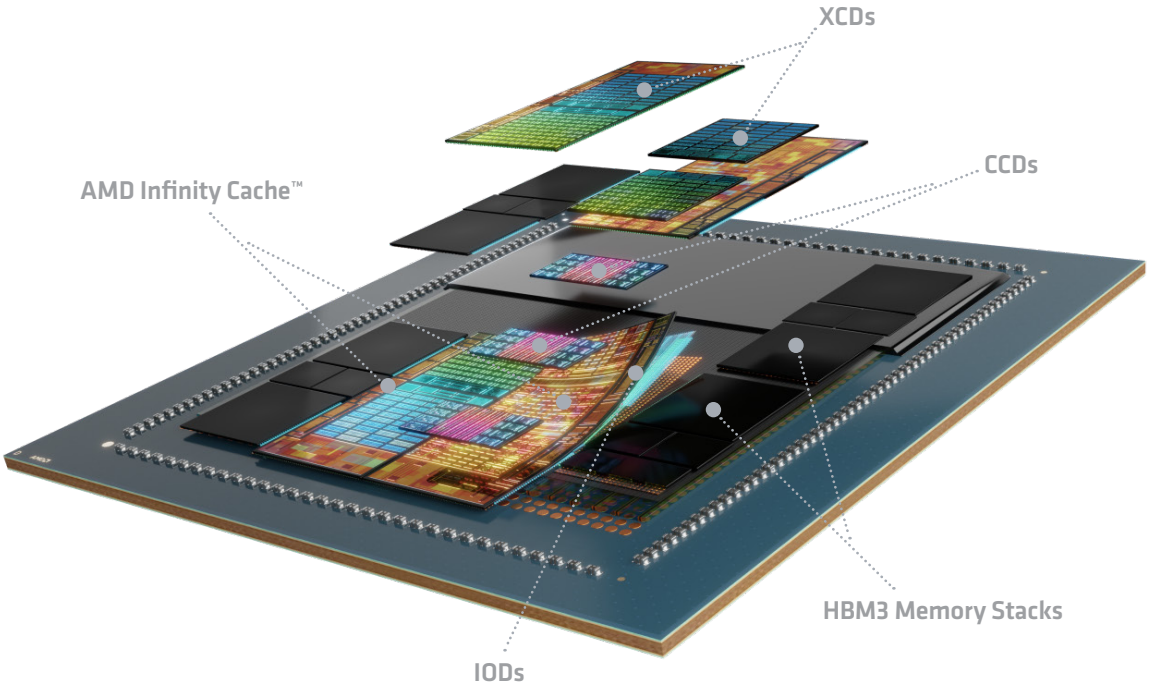
AMD has pioneered the evolution of system architecture over the last decade to unify CPU and GPU computing at an unparalleled scale. AMD Instinct™ MI250X, at the heart of the first Exascale system, was enabled by the AMD CDNA™ 2 architecture and advanced packaging, as well as AMD Infinity Fabric™, connecting the Instinct GPUs and AMD EPYC 7453s CPU with cache coherence.

The AMD Instinct™ MI200 accelerator family took initial steps towards advanced packaging, with two identical dies each incorporating the three essential elements of compute, memory, and communication. AMD CDNA™ 2 incorporated comprehensive improvements, particularly focusing on improving the communication interfaces to scale to the largest systems.

The AMD CDNA™ 3 architecture continues to lead this evolution, as AMD accelerated computing strategy embraces advanced packaging to enable heterogeneous integration. This strategy delivers outstanding performance, changing the computing paradigm with a coherent programming model that tightly couples CPUs and GPUs together to tackle the most demanding problems of our era. The AMD CDNA™ 3 architecture offers significant advancements and delivers the highest performance, efficiency, and programmability to date in the latest AMD Instinct™ MI300 Series.

As Figure 1 shows, the architecture leverages the latest advancements in 3D packaging technologies and fundamentally repartitions the compute, memory, and communication elements of the processor across a heterogenous package. The MI300 Series integrates up to 8 vertically stacked accelerator complex dies (XCD) and 4 I/O dies (IOD) containing system infrastructure, all tied together with the AMD Infinity Fabric™ technology and connecting to 8 stacks of high-bandwidth memory (HBM). At the lowest level, the computational throughput for vector and matrix data within the GPU cores is enhanced and augmented with native support for sparse data. At the macro level, this radical rethinking of the physical implementation is paired with an entirely redesigned cache and memory hierarchy that scales gracefully with the increased compute, and also integrates cache coherency as a first class citizen.

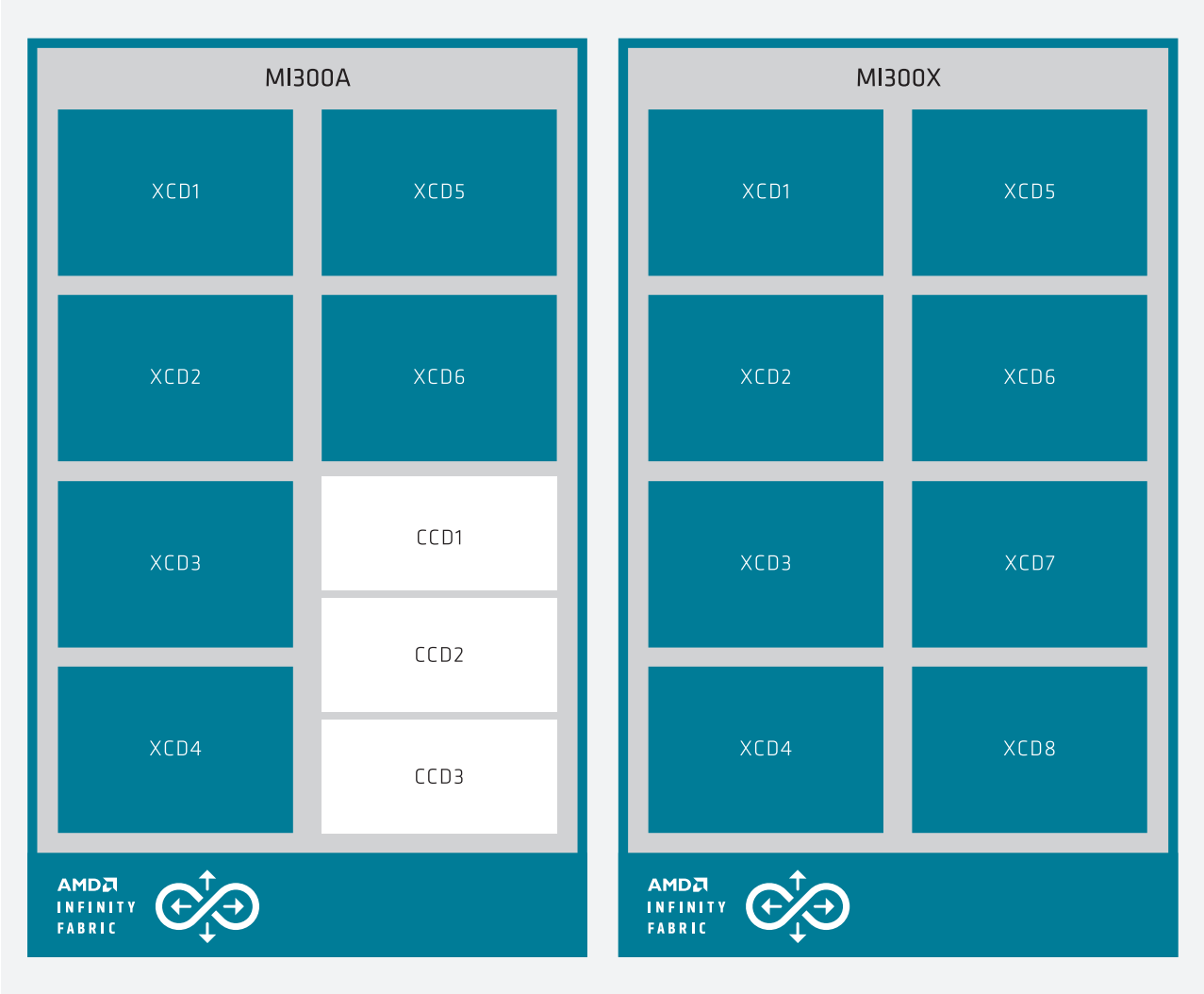
Figure 1. Advanced 3D Package and chiplet-based construction of the AMD MI300 Series processors



This design provides the versatility to construct AMD CDNA™ 3 variants, such as the MI300X discrete GPU or the MI300A APU, shown below in Figure 2. The MI300X discrete GPU focuses primarily on accelerator compute and incorporates 8 accelerator complex dies (XCD). For reduced precision data common in machine learning, the MI300X discrete GPU offers significant generational performance gains, with 3.4-6.8X the peak throughput and with peak theoretical FP8 performance of 2.6 PFLOP/s.[MI300-11] For classic HPC workloads that use single- and double-precision, the computational throughput has increased 1.7-3.4X offering 163.4 TFLOP/S FP64 Matrix for a single processor.[MI300-11]

In contrast, the MI300A APU is complete with CPU, GPU, and memory on one package. It reduces the accelerator computational capacity of the MI300X by 25% to make room for three "Zen 4" x86-based CPU dies that are tightly coupled with 6 GPU dies. The APU shares a single pool of virtual and physical memory with extraordinarily low latency. The MI300A is the world's first high-performance, data center APU, bringing tremendous ease-of-use to developers by eliminating host/device data copies and helping save substantial power and area at the system level by eliminating components such as DIMMs and CPU-to-GPU communication links.

Figure 2. Block diagram of the AMD Instinct™ MI300A APU and MI300X discrete GPU



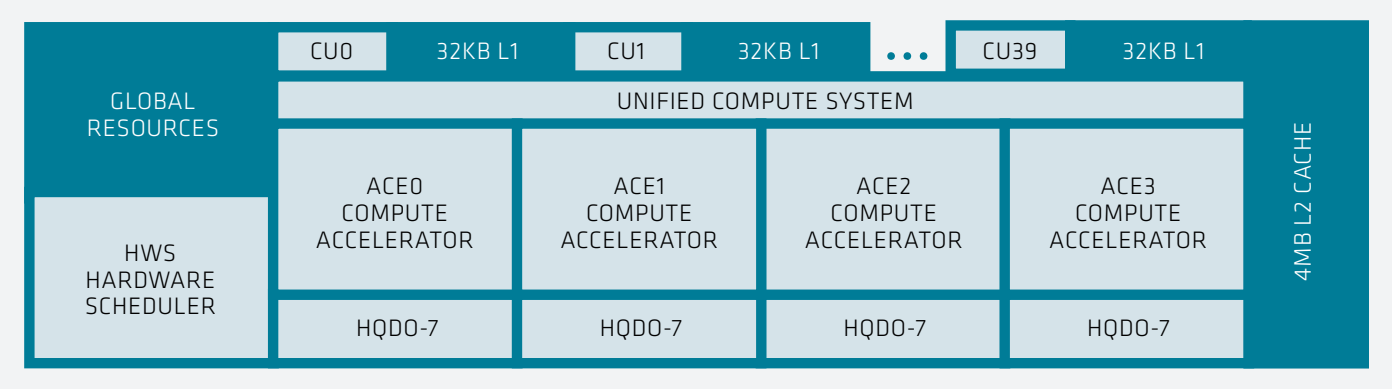
CHIPLET ARCHITECTURE REPARTITIONING

The AMD CDNA™ 2 architecture harnessed advanced packaging to couple homogeneous dies into a dual-processor package, connecting the two accelerator dies through a single high-bandwidth and low latency interconnect formed over an interposer bridge. This simple approach enabled doubling the resources by replicating compute, memory, and communication and using more silicon than would be feasible with a single die. In contrast, AMD CDNA™ 3 firmly embraces the future of chiplets. It is a single logical processor with a dozen chiplets, each specialized in both design and fabrication for compute or memory and communication, and all tied together with the AMD Infinity Fabric™ network on-chip.

AMD CDNA™ 3 COMPUTE

The first set of chiplets is the accelerator complex dies (XCDs) which contain the computational elements of the processor along with the lowest levels of the cache hierarchy, and are fabricated on TSMC's 5nm process.

Figure 3. Conceptual block diagram of an accelerator complex die (XCD)

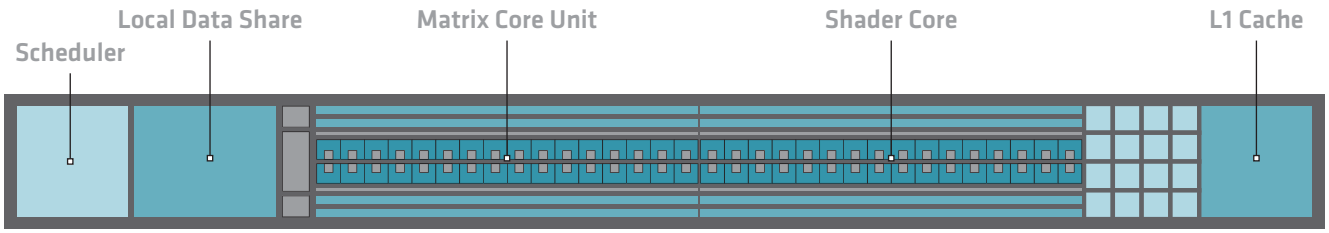


As Figure 3 illustrates above, each XCD contains a shared set of global resources such as the scheduler, hardware queues, and four Asynchronous Compute Engines (ACE) that send compute shader workgroups to the Compute Units (CUs) that are the computational heart of the AMD CDNA™ 3 architecture. The four ACEs are each associated with 40 CUs, although at the aggregate level there are only 38 CUs active, with 2 disabled for yield management. The 38 CUs all share a 4MB L2 cache that serves to coalesce all the memory traffic for the die. The AMD CDNA™ 3 XCD die is a smaller building block than the AMD Instinct MI200 Series compute die, with under half the CUs, but using more advanced packaging, the processor includes 6-8 XCDs for as many as 304 CUs total, roughly 40% more than the MI250X.[MI300-15]

AMD CDNA™ 3 COMPUTE UNIT ARCHITECTURE

As Figure 4 shows below, the AMD CDNA™ 3 compute units are complete, highly threaded and parallel processor cores including everything from instruction fetching and scheduling, execution units for scalar, vector and matrix data types, and load/store pipelines with an L1 cache and Local Data Share (LDS) that form the start of the memory hierarchy. While the compute units are architecturally similar to those in AMD CDNA™ 2, they have been comprehensively improved with major changes throughout the core to exploit greater parallelism at nearly every level and in many cases doubling or even quadrupling the performance per CU for vector and matrix workloads.

Figure 4. Conceptual block diagram of an enhanced compute unit (CU) of the AMD CDNA 3 architecture



The instruction cache is shared between two CUs and doubles the capacity from the prior generation to a 64KB and 8-way set-associative data array. This arrangement takes advantage of the reality that the overwhelming majority of the time the same instruction stream will be executed by groups of CUs, so this has the net effect of increasing the cacheable window and hit rate while keeping die area nearly constant. The AMD CDNA 3 CU generationally improves the source caching to provide better re-use and bandwidth amplification so that each vector register read can support more downstream vector or matrix operations.

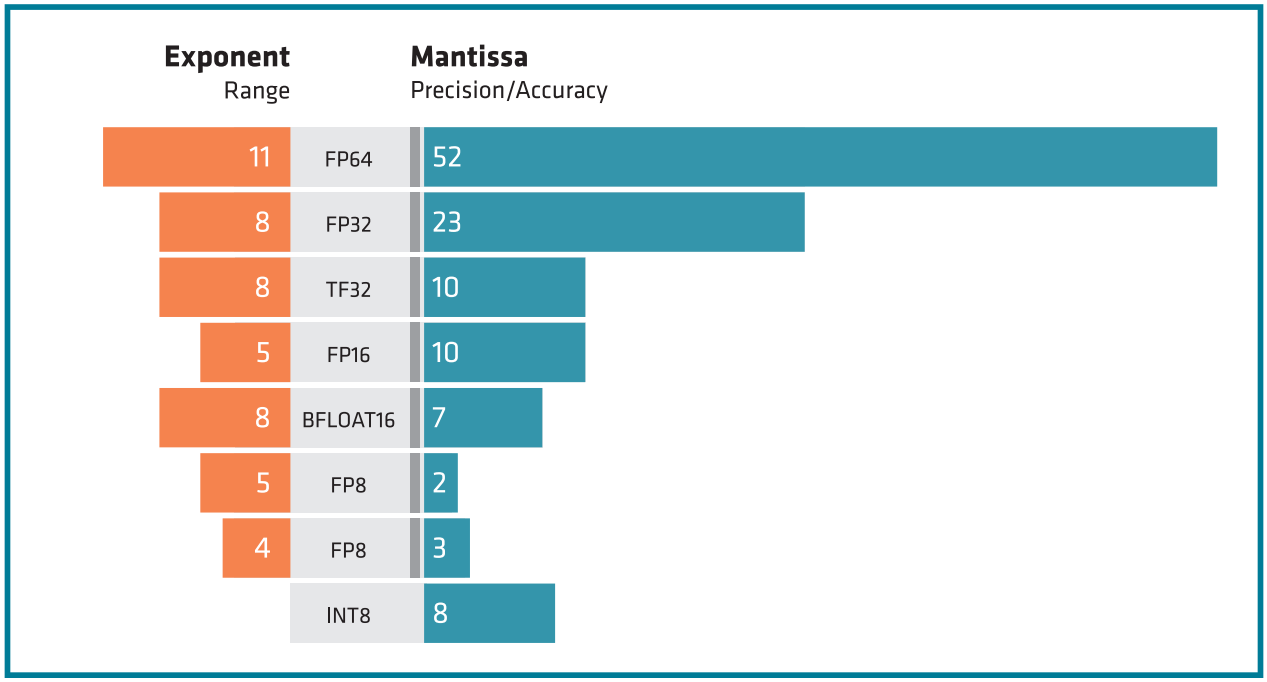
The biggest improvements in the AMD CDNA 3 CUs are in the Matrix Cores and in particular emphasizing AI and machine learning, by enhancing throughput for existing data types commonly used for cutting edge training and inference as well as adding entirely new ones. One of the greatest levers for performance in machine learning is employing more compact data types, which can save both memory and cache capacity and improve throughput and reduce power consumption. A decade ago, most machine learning applications relied on FP32, but over time the community learned to use smaller and smaller data types. For optimal training performance, the AMD CDNA 2 Matrix Cores supported FP16 and BF16, while offering INT8 for inference. As Table 1 illustrates the AMD CDNA 3 Matrix Cores triples performance for FP16 and BF16, while providing a 6.8x performance gain for INT8 compared to previous Gen MI250X accelerators.[MI300-11]

Table 1. Generational theoretical peak compute comparison for numerical formats and throughput between AMD Instinct™ MI300X discrete GPUs and MI250X GPUs.[MI300-11]

Computation	MI300 (FLOPS/clock/CU)	MI250X (FLOPS/clock/CU)	MI300X GPU (Peak TFLOP/s)	MI250X GPU (Peak TFLOP/s)	MI300 Peak Speedup
Matrix FP64	256	256	163.4	95.7	1.7x
Vector FP64	128	128	81.7	47.9	1.7x
Matrix FP32	256	256	163.4	95.7	1.7x
Vector FP32	256	128	163.4	47.9	3.4x
Matrix TF32	1024	N/A	653.7	N/A	N/A
Matrix FP16	2048	1024	1307.4	383	3.4x
Matrix BF16	2048	1024	1307.4	383	3.4x
Matrix FP8	4096	N/A	2614.9	N/A	N/A
Matrix INT8	4096	1024	2614.9	383	6.8x

To boost AI performance, the industry is constantly exploring new data types and the AMD CDNA 3 Matrix Cores adds support at full rate for the much newer FP8 and TF32 data types. TF32 is somewhat of a misnomer and it is actually a hybrid of FP16 and BF16. As Figure 5 below shows, the data format is 19-bits and combines the more precise 10-bit mantissa of FP16 with the wider range 8-bit exponent from BF16 (along with a sign bit). The name of the TF32 data type comes from the fact that for nearly all machine learning, it can easily replace FP32 without causing any accuracy problems and improve performance, whereas FP16 and BF16 have slightly different behaviors. For example, FP16 range is not as effective for recommendation and large language models, while the BF16 mantissa is too small for some vision workloads that use high-dynamic range representations.

Figure 5. Selected numerical data formats in AMD CDNA™ 3 architecture



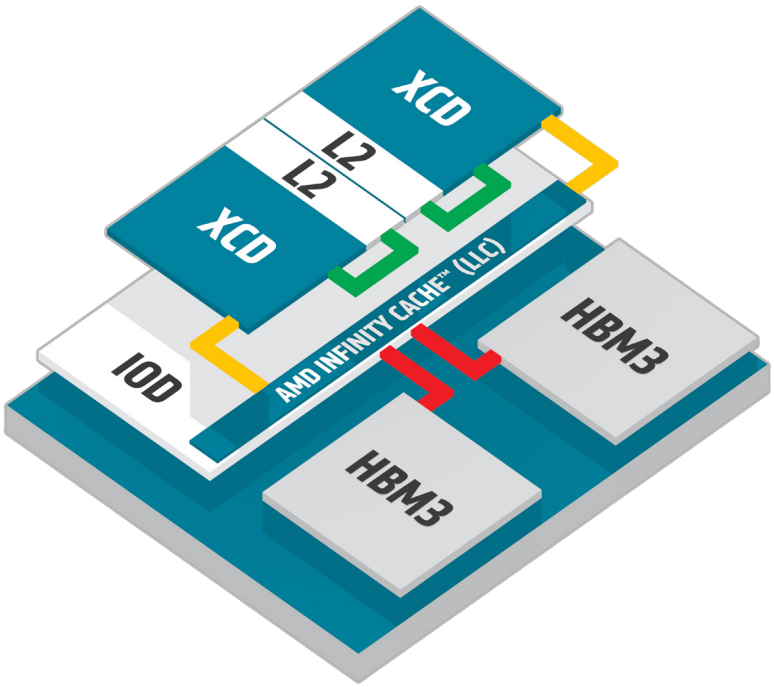
While the TF32 data type was designed for maximal ease-of-use, the FP8 data types focus purely on the smallest possible size to attain the maximum performance. The AMD CDNA™ 3 compute units support the two variants of the FP8 data type as defined in the OCP 8-bit Floating Point Specification - one with a 2-bit mantissa and a 5-bit exponent for training (E5M2) and a 3-bit mantissa with a 4-bit exponent for inference (E4M3). As Table 1 above shows, the peak theoretical computational throughput for FP8 on the MI300X discrete GPU is 16X the FP32 peak performance while the smaller datatype enables caches and memory to fit more parameters and activations, which is critical for cutting edge large language models (LLMs) that are constantly pushing the boundary of what is possible.

Last, the AMD Matrix Core Technology is now capable of efficiently handling sparse data to optimize machine learning workloads using matrix INT8, FP8, FP16, BF16. In many neural networks, a high percentage of the data will be zeroes - this is quite common in attention modules in transformer-based networks like most LLMs and also in convolution-based networks. The Matrix Core can take advantage of situations where at least two values within a group of four input values are zero (50%+ sparsity). The sparse non-zero data is represented in a compact and dense form with additional metadata tracking the locations. The dense representation of the data fits directly into the Matrix core pipeline and enables doubling the computational throughput up to an incredible 8K operations per clock for a CU.

The LDS in the AMD CDNA™ 3 compute units remain at 64KB similar to AMD CDNA™ 2 compute units. The L1 vector data cache is responsible for feeding into the vector register file and LDS and keeping the execution units fully utilized. With the significant gains in throughput for the AMD CDNA 3 Compute Units, it's no surprise that the vector data cache improved substantially to keep pace in several dimensions. The cache line size has doubled to 128B, which also doubles L1 data cache capacity in tandem to 32KB, and improves the hit rate and reduces the pressure on outer cache levels. In addition, the request bus from the data cache to the core itself and the fill path from the L2 expand to match the new line size, doubling the bandwidth to the core. One thing that remains unchanged is that the vector data cache has a very relaxed coherency model that requires explicit synchronization to offer strong coherency and ordering guarantees.

In many respects, the greatest generational changes in the AMD CDNA 3 architecture lie in the memory hierarchy outside of the Compute Units, which has been entirely re-architected to take full advantage of the heterogeneous chiplets and to enable cache coherency for co-packaged CPU chiplets in APU products. This redesign of the memory hierarchy truly starts with the shared L2 cache within the XCDs, but touches nearly every other aspect as well. The role of the L2 cache has fundamentally changed with the addition of the AMD Infinity Cache™, which is a last level cache (LLC) located on the active I/O die (IOD). Figure 6 below shows the memory architecture. Some of the more memory-oriented functions have been removed and shifted to the AMD Infinity Cache, while other aspects are new or more prominent. For example, the L2 plays a critical new role since it is the lowest level of cache where coherency is automatically maintained by the hardware. At the same time, it has been redesigned to sustain a much richer mix of resources to the CUs while isolating them from coherency traffic and optimizing the interface to the AMD Infinity Fabric™ network.

Figure 6. AMD CDNA™ 3 architecture memory architecture diagram



LEGEND

- Green line is **L2 to XCD** 51.6 TB/s (Aggregate)
- Yellow line is **LLC to XCD** 17.2 TB/s (Aggregate)
- Red line is **L2 to XCD Bandwidth** 5.3 TB/s (Aggregate)

The L2 is a 4MB and 16-way set associative cache that is massively parallel with 16 channels that are each 256KB. The L2 cache is shared by all 38 Compute Units and services requests from both the lower level instruction and data caches. On the read side each channel can read out a 128-byte cache line and the L2 cache can sustain four requests from different CUs per cycle for a combined throughput of 2KBytes/clock for each XCD. The 16 channels only support a half-line 64-byte write each with one fill request from the Infinity Fabric per clock. AMD CDNA 2 actually has 32 channels for each L2 cache, but only has at most two instances, whereas AMD CDNA 3 has collectively up to eight instances and up to 34.4 TB/s aggregate read bandwidth.

The L2 is a writeback and write-allocate design that is intended to coalesce and reduce the number of accesses that spill out and cross the AMD Infinity Fabric to the AMD Infinity Cache. The L2 itself is coherent within an XCD. The Infinity Cache includes a snoop filter covering the multiple XCD L2 caches so that the vast majority of coherent requests from other other XCDs will be resolved at the Infinity Cache without disturbing the highly utilized L2 caches.

AMD CDNA™ 3 ARCHITECTURE MEMORY

Heterogeneous integration enables the AMD CDNA 3 architecture to incorporate a substantial amount of silicon dedicated to the memory hierarchy. The IODs are manufactured on TSMC's 6nm process and vertically stacked beneath a pair of XCDs. They contain a brand new AMD Infinity Cache and the HBM3 interface to the on-package memory and are connected to the rest of the system by the AMD Infinity Fabric network.

The L2 acts as a single point of interface for each XCD, coalescing all the local memory traffic to and from the 38 CUs before it spills out to the IOD. The concept of a channel originates in the L2 (each L2 comprising 16 channels), but is critical all throughout the rest of the memory hierarchy in the IOD and beyond. Each L2 is connected to the IOD across the set of sixteen channels and each channel is 64-bytes wide, for a total of 1K-bytes per XCD at the IOD interface.

The AMD Infinity Cache is an entirely new and massive structure for the AMD CDNA 3 architecture that boosts generational performance and efficiency by increasing cache bandwidth and reducing the number of off-chip memory accesses. Typically, GPU caches are more closely aligned with and physically co-located with memory controllers and that is especially true for the AMD CDNA 3 architecture. The AMD Infinity Cache was carefully designed as a shared memory-side cache, meaning that it caches the contents of memory and cannot hold dirty data evicted from a lower level cache. This has two significant benefits. First, the AMD Infinity Cache doesn't participate in coherency and does not have to absorb or handle any snoop traffic, which significantly improves efficiency and reduces the latency of snooping from lower level caches. Second, it can actually hold nominally uncacheable memory such as buffers for I/O.

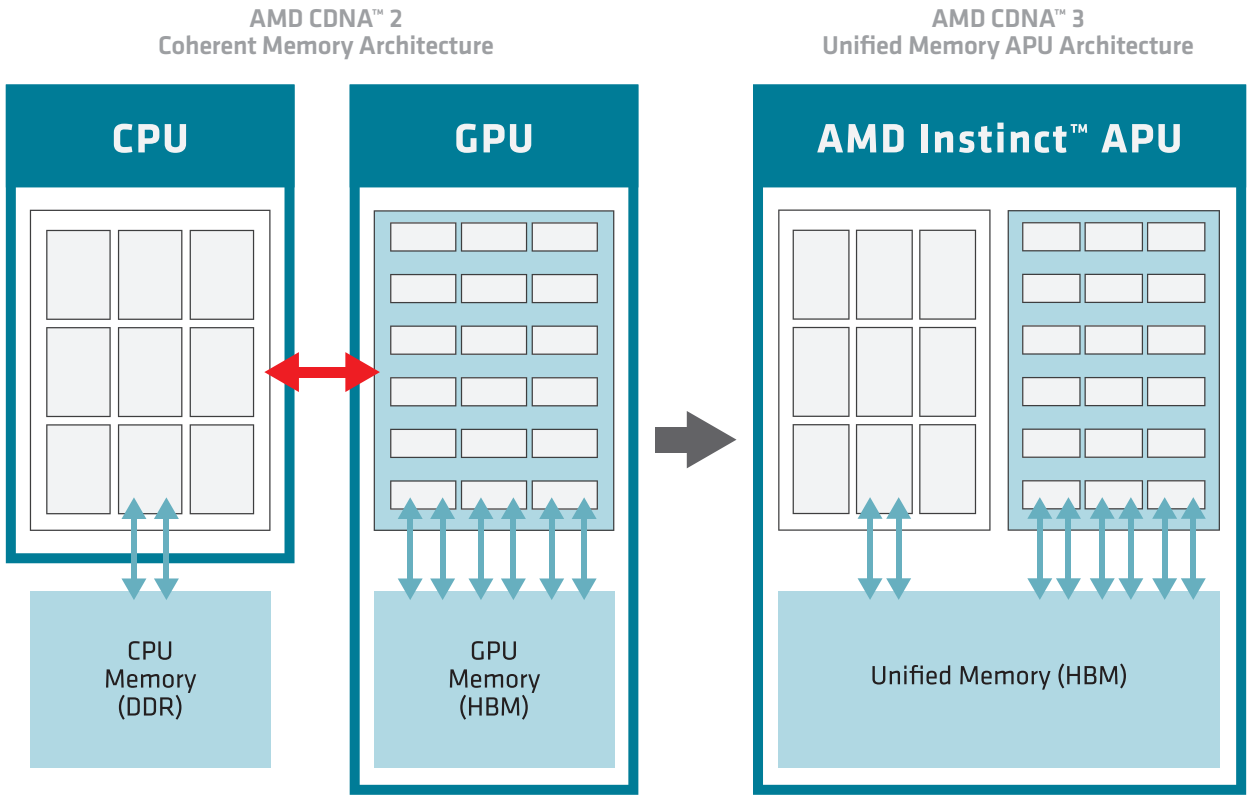
Just like the L2 cache, the AMD Infinity Cache is 16-way set-associative, and it is built around the concept of channels. Each stack of HBM memory is associated with 16 parallel channels. A channel is 64-bytes wide and connects to 2 MB of data arrays that are banked to sustain simultaneous reads and writes. In total, there are eight stacks of HBM across the four IODs, for 128 channels or 256MB of data. The peak bandwidth from the Infinity Cache is an astounding 17.2 TB/s, which is nearly as much as the total from the previous generation L2 caches and a welcome addition to the overall memory hierarchy.

Moving beyond the AMD Infinity Cache, each IOD fans out through the package to two stacks of memory. The AMD CDNA 3 architecture upgrades the interface from HBM2e to the latest HBM3. The memory controllers drive a bus that operates at 5.2 Gbps and each stack contains 16GB or 24GB of memory. Collectively, the HBM3 memory is 128GB on MI300A and 192GB on MI300X per accelerator and both have an astounding 5.3 TB/s peak theoretical memory bandwidth.[MI300-13, MI300-14]

Additionally, in the MI300A the HBM3 memory is unified and shared between the GPU and CPU, dramatically reducing the latency and improving communication throughput. As Figure 7 below illustrates, this also pays dividends at the system-level by eliminating the interconnect and also the DDR memory, saving space and power.

Figure 7. Unified memory architecture of the AMD CDNA™ 3 APU compared to the AMD CDNA™ 2 coherent memory architecture

AMD Instinct™ MI300A: World's first Data Center APU Engineered for Next-Level HPC

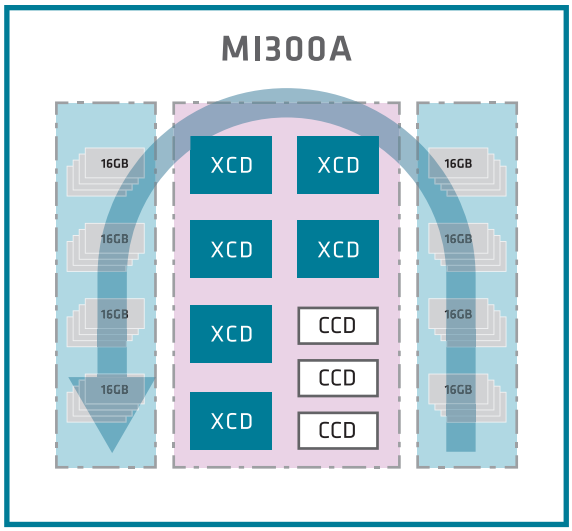


The AMD Instinct MI300 Series family of GPUs include 6-8 XCDs that can be spatially partitioned to appear as multiple virtual GPUs. In the simplest case as seen in Figure 8 below, the GPU is presented as a single processor so that all XCDs operate for a single user or problem. At the other end of the spectrum, the GPU can be divided into as many partitions as XCDs, with each running independently. In this example, each XCD could operate on a separate stream of input queries for inference and offer concurrency with greater isolation. The AMD Instinct MI300X discrete GPU can be configured with up to eight partitions, one per XCD, while the AMD Instinct MI300A APU has a maximum of three partitions with two XCDs per partition. AMD Instinct MI300 Series family of GPUs supports Single Root IO Virtualization (SR-IOV) that provides isolation of Virtual Functions (VF's) and protect a VF from accessing information or state of the Physical Function (PF) or another VF. In addition to GPU partitons, the HBM can also be partitioned on the MI300X discrete GPU. Known as NUMA partitions per socket or NPS, the HBM can be one or four partitions and is a separately configured from the GPU partitioning. The memory partitioning must be equal to or smaller than the number of GPU partitions. Example: NPS4 memory partitioning can be combined with four or eight GPU partitions.

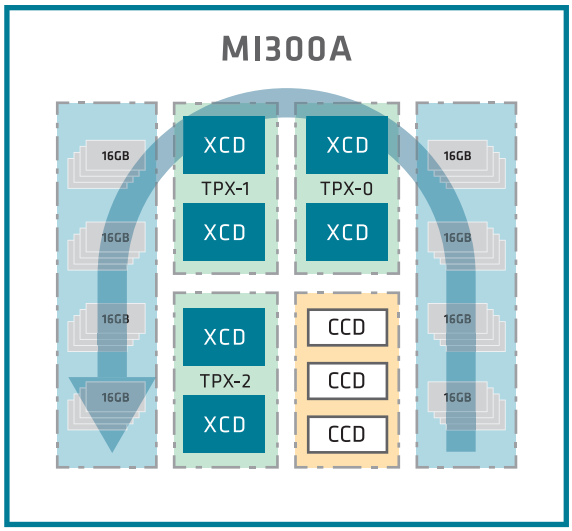
Figure 8. AMD Instinct™ MI300 Series spacial partitioning diagram for virtualized environments

AMD Instinct™ MI300A: Partitioning Examples

HBM unified memory space, partitioning, and socket physical memory map are shown in the examples below.



(a)
Single Partition: SPX, NPS 1



(b)
Triple Partition: TPX, NPS 1

CPU & ACCELERATOR MEMORY

- 128 GiB of DRAM
- Interleaved between HBM stacks: switch stack every 4KiB through physical memory space
- All allocations controlled by the OS: accelerator driver is a client of the OS memory manager

LEGEND

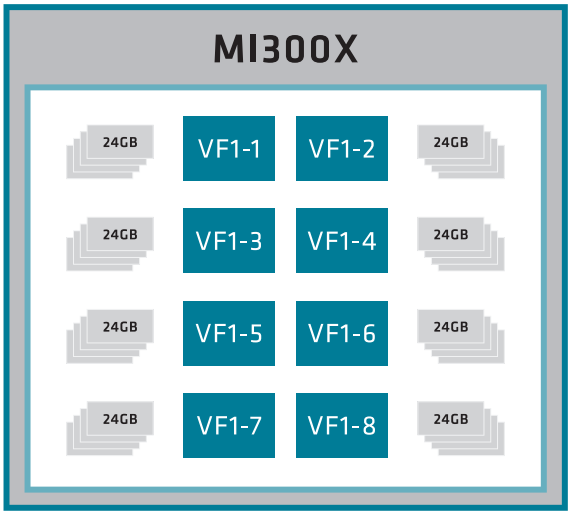
- MI300A CCD
- MI300X XCD
- NPS1 Partition
- SPX Compute and CPU Partition
- TPX Compute Partition (TPX-0, TPX-1, TPX-2)
- CPU Partition
- HBM Unified Memory Space

SOCKET PHYSICAL MEMORY MAP

Low address High address

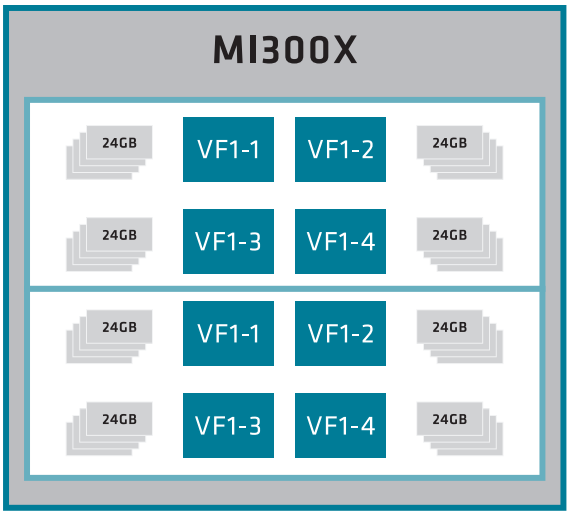
AMD INSTINCT

AMD Instinct™ MI300X: Partitioning and Virtualization Examples



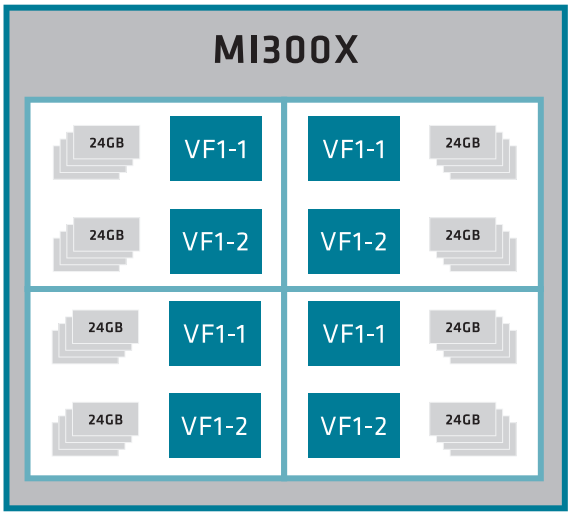
(a)
Single Partition: SPX, NPS 1

8 XCDs per Partition / 8 XCDs per VM,
192 GB per Partition / 192 GB per VM



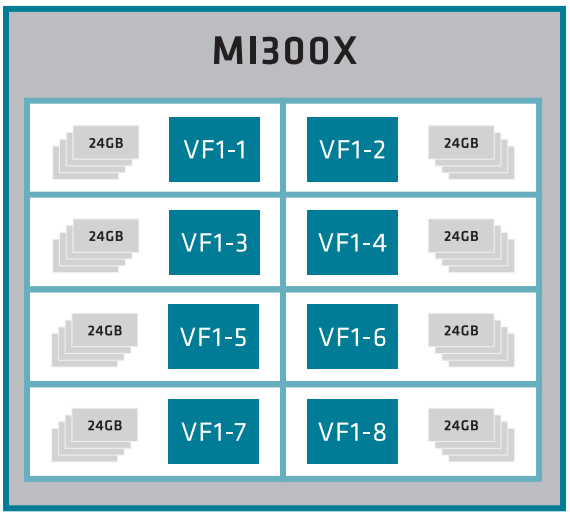
(b)
Two Partitions: DPX, NPS 1

4 XCDs per Partition / 4 XCDs per VM,
96 GB per Partition / 96 GB per VM



(c)
Four Partitions: QPX, NPS 1 & 4

2 XCDs per Partition / 2 XCDs per VM,
48 GB per Partition / 48 GB per VM



(a)
Eight Partitions: CPX, NPS 1

1 XCDs per Partition / 1 XCDs per VM,
24 GB per Partition / 24 GB per VM

LEGEND

- MI300X XCD / Virtual Function (e.g. 'VF1')
- Partition / Virtual Machine

COMMUNICATION AND SCALING

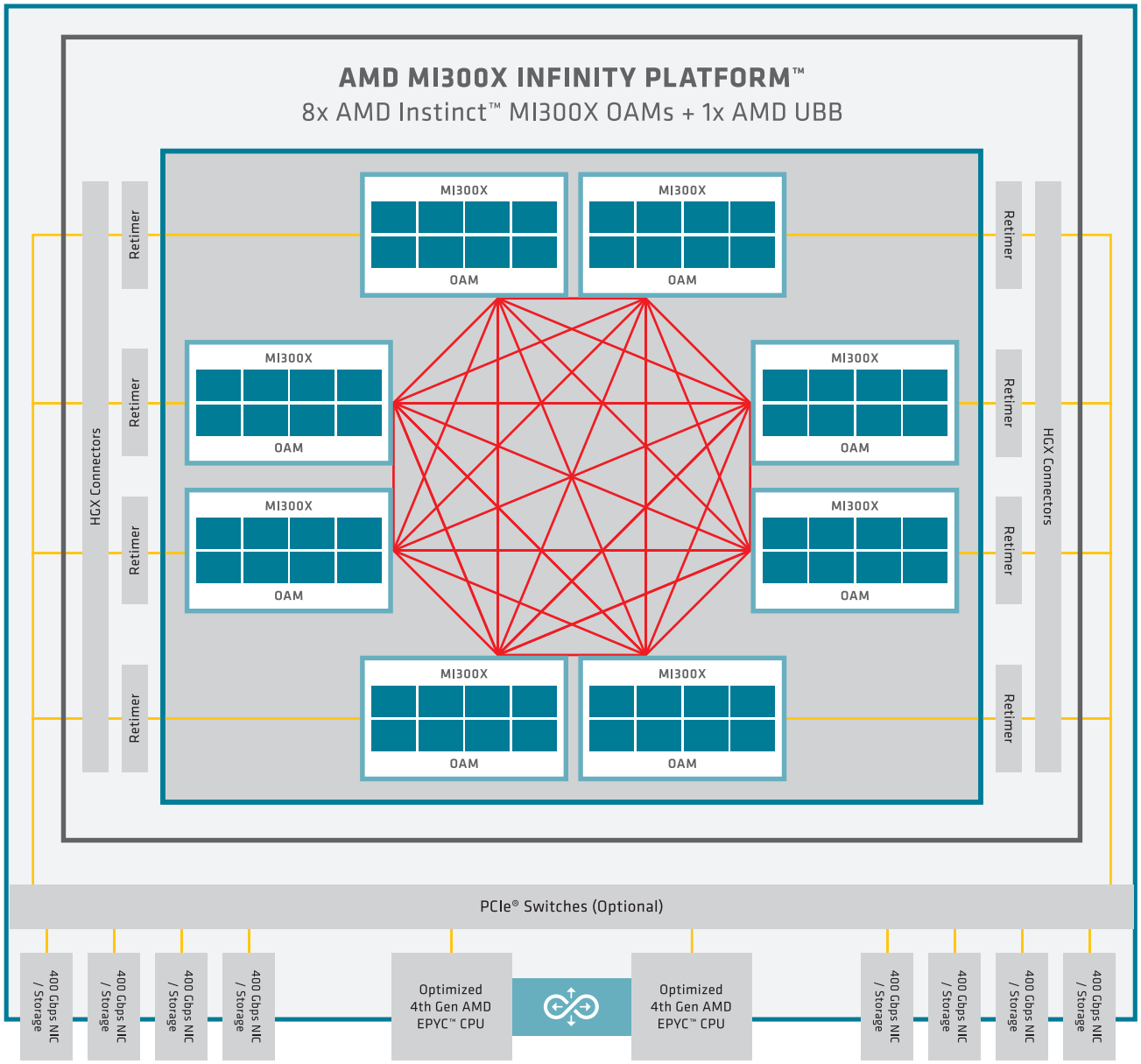
To build the world's largest exascale-class supercomputers and AI systems that can advance the state-of-the-art, a single processor is never enough. Instead architects must focus on the entire system – from the chip to the platform design to the rack and then to data center level. The AMD CDNA 2 architecture was a dramatic leap forward in capabilities adopting the 3rd Gen AMD Infinity architecture, including AMD Infinity Fabric™ technologies within package, between packages, and to host processors. The AMD CDNA 3 architecture takes communication and scaling to the next level using the 4th Gen Infinity architecture fabric more prolifically inside the package and enhances the efficiency and performance across the board. However, the heterogeneous integration for the AMD CDNA 3 family gives AMD the unique opportunity to push scalability in two different directions with the AMD Instinct MI300X discrete GPU and the AMD Instinct MI300A APU.

For AMD CDNA 3, the communication links have been systematically enhanced to operate at up to 32Gbps and redistributed across the IODs. Each IOD includes two 16-lane, bi-directional inter-package AMD Infinity Fabric links for connecting to other AMD accelerators. One of the links is multi-purpose and can be configured to act as a x16 PCIe® Gen 5 for pure I/O functionality.

Figure 9. AMD Instinct™ MI300X discrete GPU platform 8-socket GPU design

AMD MI300X Infinity Platform™: 8 OAM + AMD UBB Node Example

The AMD MI300X Infinity Platform™ provides an efficient form factor for the flagship AI training and inference topology with an 8-GPU MI300X node as illustrated below.



Light blue is AMD Infinity Fabric™ bi-directional CPU to CPU link

Yellow lines are PCIe® Gen5

Red lines are AMD Infinity Fabric™bi-directional links

MI300X XCD

As Figure 9 illustrates above, the MI300X discrete GPU uses these seven high-bandwidth and low-latency AMD Infinity Fabric links to form a fully connected 8-GPU system. Each GPU is also connected to the host CPU via a x16 PCIe® Gen 5 link. This approach is commonly adopted using the OCP Universal Base Board (UBB) form factor, which builds on a variety of industry-standard technologies to enable easy to build and deploy systems. Compared to the prior generation, this 8-GPU node is intrinsically faster and more efficient for communication patterns such as allreduce and allgather which are employed in gradient summation and data parallel sharding for machine learning.

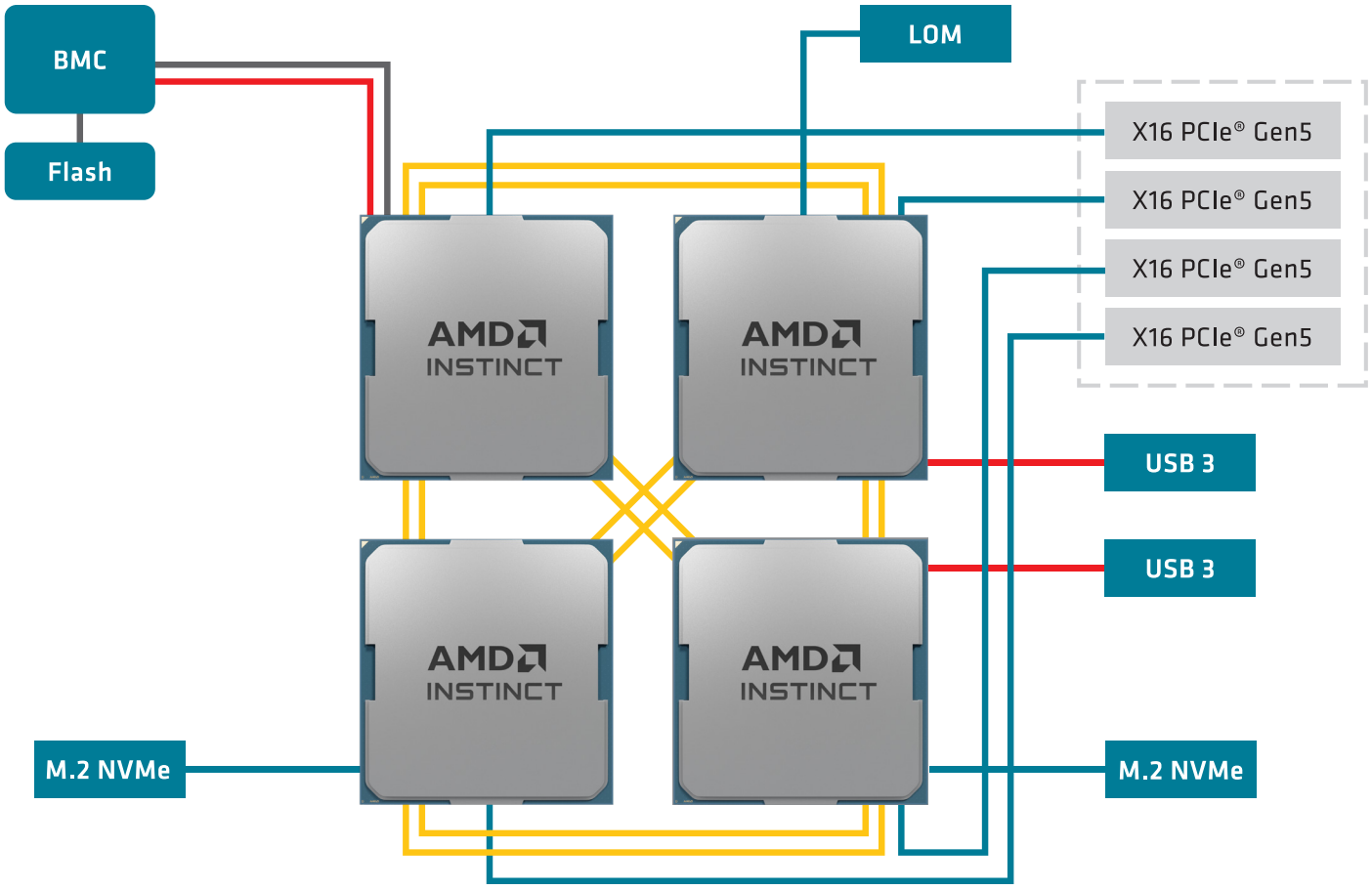
For the MI300A APU, the on-package integration of CPU cores and unified memory is even more transformative. In the previous generation, the AMD EPYC™ processor and the MI250X GPU connected via two AMD Infinity Fabric links with 144GB/s throughput and package-to-package latency between the two devices and their associated memories. On the MI300A APU, the on-package AMD Infinity Fabric connects both the accelerator complex dies (XCDs), and the CPU complex dies (CCDs) directly into the shared Infinity Cache and 8-stack of HBM3 at chiplet latency and interface throughput.

Turning to the node level, the MI300A APU also delivers greater fabric bandwidth between processors than the prior generation. Many HPC systems focus on 4-processor nodes where, as Figure 10 below shows, each processor is fully connected to its peers using two AMD Infinity Fabric links with 256GB/s of bandwidth.

Figure 10. Typical AMD Instinct™ MI300A APU platform design with 4 APUs connected with AMD Infinity Fabric™ links

AMD Instinct™ MI300A: 4-APU Node Example

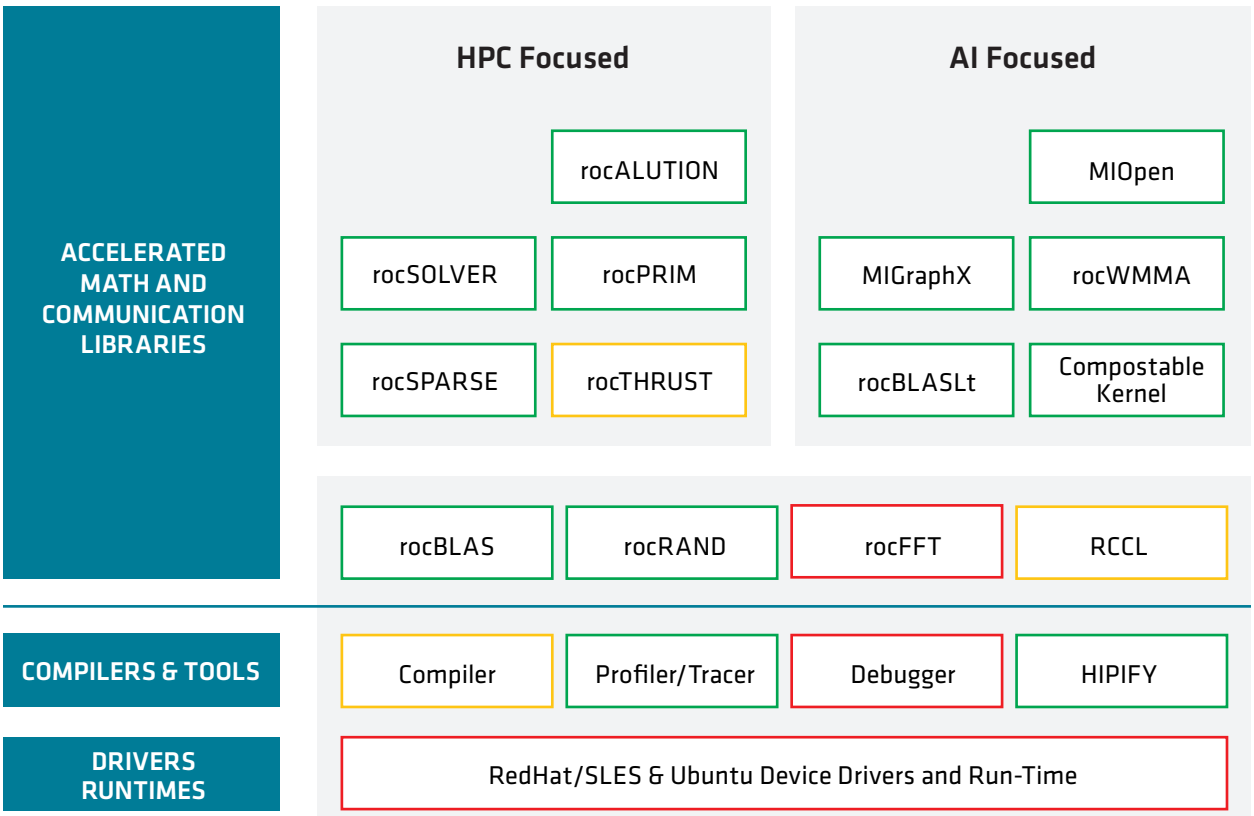
Flagship HPC Topology with MI300A



AMD ROCM™ SOFTWARE STACK FOR AMD INSTINCT ACCELERATORS

For accelerated computing and especially at scale, software is absolutely critical to the success of the system. A robust software stack and healthy ecosystem will unlock the creativity and capabilities of software developers and customers. As Figure 11 below illustrates, the AMD ROCm™ software strategy is built on open-source, which empowers the community by giving developers and customers visibility into the stack and also the ability to improve and adapt it to their needs.

Figure 11. The AMD ROCm™ open software platform for accelerated computing is a comprehensive set of open-source APIs, compilers, libraries, and development tools



LEGEND

- Green lines are MIT/BSD License
- Yellow lines are Apache License
- Red lines are GPL License

The AMD open-source ROCm software platform empowers the accelerated computing community to innovate on top of a robust, flexible stack designed for scalability. The ROCm ecosystem consists of modular, high-performance frameworks, libraries and compilers. These components work in concert to extract the full potential of heterogeneous architectures. The platform's open-source philosophy gives developers complete visibility while enabling customization and co-development. Users can optimize ROCm software platform's runtimes, programming models and utilities based on their workloads and scale requirements.

ROCm provides low friction portability, simplified programmability and usability to accelerate time to results with drop-in support for leading frameworks. The extensive ROCm library of performant software tools, hardware abstraction and cross-platform capabilities boost productivity and assist in achieving maximum performance across environments. A simplified programming model across heterogeneous infrastructure helps speed up development cycles and enables to go from prototype to production quickly.

The AMD growing repository of optimized applications is powered by close collaborations with leading HPC and AI industry partners. The ROCm open software platform and HIP cross-platform programming model foster innovation through open design and development. AMD continuously participates in and contributes to multi-hardware open-source projects like OpenXLA, Triton, and MLIR to shape the future of performance portability standards. Ongoing investment in ecosystem initiatives aims to make AI faster, more scalable, and flexible across computing environments.

The building block for AMD ROCm is the Heterogeneous-computing Interface for Portability (HIP) runtime, API, and language. The HIP language is similar to C++ and as the name implies, the toolchain is designed so that a single codebase using HIP will produce high-performance code for GPUs from AMD and other companies. The philosophy behind HIP is that both portability and performance come through a combination of highly tuned libraries and code generators that make writing or converting heterogeneous applications easy.

The AMD ROCm stack, shown in Figure 11 above, consists of a collection of drivers, development tools, and APIs that enable GPU programming from low-level kernels to end-user applications. Figure 12 below shows major libraries in the AMD ROCm open ecosystem, including fundamental computing motifs such as sparse matrices, iterative solvers, random number generation, and fast-Fourier transforms. As one example, MIOpen is the AMD library for high performance machine learning primitives that provides GPU-optimized implementations of common operations used in neural nets such as convolution, batch normalization, activation functions, pooling, and more. Using the HIP interface to these libraries helps ensure expected performance for AMD Instinct accelerators.

Figure 12. AMD ROCm™ software Core Libraries - Source Code Information

LIBRARY	DESCRIPTION	SOURCE
CK	Programming model for performance-critical kernels for ML workloads	https://github.com/ROCmSoftwarePlatform/composable_kernel
rocBLAS	Basic Linear Algebra Subroutines	https://github.com/ROCmSoftwarePlatform/rocBLAS
rocFFT	Fast Fourier Transfer Library	https://github.com/ROCmSoftwarePlatform/rocFFT
rocPRIM	Low Level Optimized Parallel Primitives	https://github.com/ROCmSoftwarePlatform/rocPRIM
rocRAND	Random Number Generator Library	https://github.com/ROCmSoftwarePlatform/rocRAND
rocSOLVER	Implementation of Lapack Routines	https://github.com/ROCmSoftwarePlatform/rocSOLVER
hipSPARSELt	Basic Linear Algebra Subroutines for sparse computation	https://github.com/ROCmSoftwarePlatform/hipSPARSELt/
rocThrust	C++ parallel algorithms library	https://github.com/ROCmSoftwarePlatform/rocThrust
MIOpen	Deep learning Solver Library	https://github.com/ROCmSoftwarePlatform/MIOpen
RCCL	Communications Primitives Library based on the MPI equivalents	https://github.com/ROCmSoftwarePlatform/rccl
MIGraphX	Graph inference engine that accelerates ML model inference	https://github.com/ROCmSoftwarePlatform/AMDMIGraphX

To learn more about ROCm source code and documentation, please refer to Figure 13 below.

Figure 13. AMD ROCm™ Source Code & Documentation

ROCm™ Source Code	ROCm CoreTechnology and Documentation	Contains low-level (driver, system management, etc) components and the base repository of the ROCm project	https://github.com/RadeonOpenCompute
	Basic Linear Algebra Subroutines	Contains developer tools (HIPIFY, rocprofiler, rocgdb, etc.) and programming language repositories, including HIP	https://github.com/ROCm-Developer-Tools
	ROCm Software Platform Repository	Contains computational, communication, and AI libraries.	https://github.com/ROCmSoftwarePlatform
	GPU Open-Professional Compute Libraries	Contains computer vision respositories	https://github.com/GPUOpen-ProfessionalCompute-Libraries
ROCm™ Documentation	ROCm	AMD ROCm Documentation Hub	https://rocm.docs.amd.com/en/latest/
	AMD Developer Central	Resources to develop using AMD Products	https://www.amd.com/en/developer.html

AMD Instinct GPUs offer integrated Support for Machine Learning and Deep Learning frameworks such as PyTorch, TensorFlow, and JAX. These are optimized to leverage AMD Instinct GPU hardware extensions (MIOpen, ROCm libs) and offer interoperability with ONNX Runtime for optimization across backends.

AMD accelerates machine learning workloads on GPUs through extensive optimizations applied at multiple levels:

- Low-Level Libraries: MIOpen, rocBLAS and other AMD created libraries are finely tuned to boost performance on individual operators and primitives leveraged by ML frameworks like TensorFlow and PyTorch.
- Intermediate Representations: Compiler-based intermediate representations (IRs) like OpenXLA and Triton apply graph and node level optimizations on computational graphs to tune performance for AMD hardware. This enables cross-platform portability.

These multi-tier, end-to-end optimizations spanning operators, models and hardware, provide performant, flexible and future-ready machine learning acceleration on AMD GPUs.

AMD ROCm 6 AND ITS KEY FEATURES ENABLED FOR AMD CDNA 3 ARCHITECTURE BASED ACCELERATORS:

With the introduction of ROCm 6, AMD has enabled the following features for the latest AMD Instinct MI300 Series accelerators:

1. Optimized performance: new and improved FP64 matrix operations and advanced cache handling, along with improved kernel launch latency and runtime
2. ROCm 6 brings support for TF32 and FP8 Data types on AMD Instinct MI300 Series GPUs
3. ROCm 6 introduces a new hipSparseLT Library, which enables the use of sparsity cores
4. Enabling developer success: prepackaged HPC and AI/ML frameworks ready for download on the AMD Infinity Hub; streamlined and improved tools.

For more information on the AMD ROCm open software platform and AMD Instinct supporting software, visit AMD.com/ROCm.

Below in Table 2, AMD Instinct MI300 Series accelerator product specifications and features are provided.

AMD INSTINCT™ MI300 SERIES PRODUCT OFFERINGS

Table 2. AMD Instinct™ MI300 Series accelerators specifications and features

	M1300A APU	M1300X DISCRETE GPU
ARCHITECTURE	**AMD CDNA 3**	**AMD CDNA 3**
ACCELERATED COMPLEX DIES (XCD)	6	8
COMPUTE UNITS	228	304
STREAM PROCESSORS	14,592	19,456
MATRIX CORES	912	1,216
MAX ENGINE CLOCK (PEAK)	2,100 MHz	2,100 MHz
AMD “ZEN 4” CPU CHIPLETS (CCD)	3	NA
TOTAL “ZEN 4” X86 CORES	24	NA

AMD INSTINCT

	M1300A APU	M1300X DISCRETE GPU
ARCHITECTURE	**AMD CDNA 3**	**AMD CDNA 3**
UNIFIED MEMORY ACROSS CPU & GPU	Yes	NA
TRANSISTOR COUNT	146 Billion	153 Billion
PERFORMANCE (PEAK THEORETICAL)		
FP64 VECTOR	61.3 TF	81.7 TF
FP32 VECTOR	122.6 TF	163.4 TF
FP64 MATRIX	122.6 TF	163.4 TF
FP32 MATRIX	122.6 TF	163.4 TF
TF32 MATRIX | TF32 (SPARSITY)	490.3 TF | 980.6 TF	653.7 TF | 1,307.4 TF
FP16 | FP16 (SPARSITY)	980.6 TF | 1,961.2	1,307.4 TF | 2,614.9 TF
BF16 | BF16 (SPARSITY)	980.6 TF | 1,961.2	1,307.4 TF | 2,614.9 TF
FP8 | FP8 (SPARSITY)	1,961.2 TF | 3,922.3 TF	2,614.9 TF | 5,229.8 TF
INT8 | INT8 (SPARSITY)	1,961.2 TOPs | 3,922.3 TOPs	2,614.9 TOPs | 5,229.8 TOPs
MEMORY		
MEMORY CAPACITY	128GB HBM3	192 GB HBM3
MEMORY INTERFACE	1024-bits x 8 Stacks HBM3	1024-bits x 8 Stacks HBM3
MEMORY CLOCK	5.2 GT/s	5.2 GT/s
MEMORY BANDWIDTH (PEAK)	up to 5.3 TB/sec	up to 5.3 TB/sec
L1 CACHE	32 KiB	32 KiB
L2 CACHE	4 MB	4MB
AMD INFINITY CACHE™	256 MB	256MB

	M1300A APU	M1300X DISCRETE GPU
SCALE-UP | SCALE-OUT		
DEVICE I/O CONNECTIONS	4x16 AMD Infinity Fabric™ Links 4x16 PCIe® Gen 5 or AMD Infinity Fabric™ Links	7 x16 AMD Infinity Fabric™ links 1x16 PCIe® Gen 5 to host CPU
P2P RING PEAK AGGREGATE I/O BANDWIDTH	384 GB/S (4 APUs)	896 GB/s (8 GPUs)
TOTAL PEAK AGGREGATE I/O BANDWIDTH	1024 GB/s	1024 GB/s
BUS INTERFACE	PCIe® Gen 5 Support	PCIe® Gen 5 Support
VIRTUALIZATION		
SR-IOV SUPPORT	Yes	Yes
PARTITIONS	Up to 3	Up to 8
DECODERS	HVEC/H.265, AVC/H.264, V1, or AV1	HVEC/H.265, AVC/H.264, V1, or AV1
RAS FEATURES		
FULL-CHIP ECC	Yes	Yes
PAGE RETIREMENT	Yes	Yes
PAGE AVOIDANCE	Yes	Yes
BOARD DESIGN | PACKAGING		
FORM FACTOR	SH5 Socket	OAM
THERMAL	Passive and Liquid	Passive and Liquid
MAX POWER	550W or 760W	750W

CONCLUSION

The AMD CDNA™3 architecture builds on the previous generation exascale-class architecture, and takes advantage of heterogeneous integration to unify traditional and accelerated computing, delivering dramatic improvements in performance and efficiency. The advanced 3D packaging enables the repartitioning and optimization of compute, memory, and communication resources across a dozen or more specialized chiplets all connected with AMD Infinity Fabric™ technology to produce the innovative AMD Instinct MI300 Series accelerators.

The heterogeneous packaging of the MI300A APU enables combining AMD "Zen 4" x86-based CPU chiplets with AMD CDNA 3 XCD chiplets into a single APU computing platform with full cache coherency and unified memory between the two. This change requires rethinking many memory and communication aspects of the overall architecture. But a unified architecture is absolutely groundbreaking in the data center and offers innovative opportunities for developers and customers to solve new and exciting computational problems..

The 3D packaging is also critical for boosting performance across the board. For example, by moving to many XCDs dedicated specifically to compute, the MI300X discrete GPU offers up to 6.8X the peak theoretical throughput of the previous generation: 163.4 TFLOP/s for both single (FP32)- and double-precision (FP64 Matrix) and over 2.5 PFLOP/s 8-bit precision (FP8) for machine learning.[MI300-11] At the same time, the AMD CDNA 3 architecture incorporates new features like efficient handling of 4:2 sparse data and FP8 numerical formats for machine learning to further boost efficiency and performance.

The AMD CDNA 3 memory hierarchy and communication was fundamentally transformed by the repartitioning and migrated into the new I/O dies (IODs) with efficient coherency as key goal. The dedicated silicon incorporates a massive new 256MB AMD Infinity Cache™ that is shared by both traditional and accelerated computing elements, boosting bandwidth and enabling easy data sharing. This new cache in turn allows for optimizing the lower levels of the cache hierarchy for dedicated large local bandwidth. The external memory is also coherently shared between both traditional and accelerated computing elements and is upgraded to a faster HBM3 interface with 5.3TB/s of peak theoretical bandwidth and up to 192GB of capacity on the MI300X discrete GPU.

This heterogenous repartitioning places new demands on communication both within the package and between processors in a system. AMD Infinity Fabric technology is the key to the advanced 3D packaging as it ties together all the components over a massive 4TB/s fabric, enabling low-latency and coherent communication between accelerated compute elements, traditional compute elements, and the broader memory hierarchy such as the AMD Infinity Cache and HBM3. At the system level, the external interfaces are generationally faster and also enable building fully connected 4-APU and 8-GPU platforms that can handle key communication algorithms more efficiently.

Collectively, the AMD CDNA 3 architectural improvements give AMD the ability to create the AMD Instinct MI300 Series family of products, which push innovation in several directions. The dedicated MI300X discrete GPU is tailored for maximum compute and extreme machine learning workloads, while the MI300A APU is the first data center APU and creates an incredibly efficient system-level architecture for HPC. This breadth of the family is a boon for customers who can rely on AMD to push the frontiers of computing and find the right solution for the most critical emerging computing problems and opportunities.

For more information on the AMD Instinct MI300 Series of products, visit AMD.com/INSTINCT.

ENDNOTES

MI300-11- Measurements conducted by AMD Performance Labs as of November 11th, 2023 on the AMD Instinct™ MI300X (750W) 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 (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), 1307.4 TFLOPS peak theoretical Bfloat16 format precision (BF16), 2614.9 TFLOPS peak theoretical 8-bit precision (FP8), 2614.9 TOPs INT8 floating-point performance.

The results calculated for the AMD Instinct™ MI250X (560W) 128GB HBM2e OAM accelerator designed with AMD CDNA™ 2 5nm FinFET process technology 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), TF32 (N/A), 383.0 TFLOPS peak theoretical half precision (FP16), 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), FP8 (N/A), 383.0 TOPs INT8 floating-point performance. * MI200 Series GPUs don't support TF32, FP8 or sparsity

MI300-13: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300X memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8).

MI300-14: Calculations conducted by AMD Performance Labs as of November 7, 2023, for the AMD Instinct™ MI300A APU accelerator 760W (128 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 5.325 TFLOPS peak theoretical memory bandwidth performance. MI300A memory bus interface is 8,192 (1024 bits x 8 die) and memory data rate is 5.2 Gbps for total peak memory bandwidth of 5.325 TB/s (8,192 bits memory bus interface * 5.2 Gbps memory data rate/8). The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.2 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Server manufacturers may vary configuration offerings yielding different results.

The AMD Instinct™ MI250 (500W) / MI250X (560W) OAM accelerators (128 GB HBM2e) designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 128 GB HBM3 memory capacity and 3.277 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 8,192 (4,096 bits times 2 die) and memory data rate is 3.20 Gbps for total memory bandwidth of 3.277 TB/s ((3.20 Gbps*(4,096 bits*2))/8).

MI300-15: The AMD Instinct™ MI300X (750W) accelerator has 304 compute units (CUs), 19,456 stream cores, and 1,216 Matrix cores. The AMD Instinct™ MI250 (560W) accelerators have 208 compute units (CUs), 13,312 stream cores, and 832 Matrix cores. The AMD Instinct™ MI250X (500W/560W) accelerators have 220 compute units (CUs), 14,080 stream cores, and 880 Matrix cores.