11.1 AMD Instinct™ MI300 Series Modular Chiplet Package – HPC and AI Accelerator for Exa-Class Systems

Alan Smith[1], Eric Chapman[1], Chintan Patel[1], Raja Swaminathan[1], John Wuu[2], Tyrone Huang[3], Wonjun Jung[3], Alexander Kaganov[3], Hugh McIntyre[4], Ramon Mangaser[5]

[1]AMD, Austin, TX; [2]AMD, Fort Collins,CO; [3]AMD, Markham, Canada
[4]AMD, Santa Clara, CA; [5]AMD, Boxborough, MA

2024 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3503-0620-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSCC49657.2024.10454441

The AMD Instinct™ MI300 Series accelerators were conceptualized to extract maximum HPC and AI capability from the latest silicon and advanced packaging technology, designed to operate as CPU hosted PCIe® device, MI300X, as well as a self-hosted accelerated processing unit (APU), MI300A. AMD chiplet capabilities and advanced packaging allow AMD's first-ever integration of data center class CPU, GPU accelerated compute, AMD Infinity Cache, and 8-stack HBM3 memory system into a single package. Observing that many AI and HPC operators are memory bound, AMD targeted MI300 to deliver over 5TBps of HBM3 peak bandwidth.

MI300X (Fig. 11.1.8) targets traditional 2P CPU servers hosting eight GPU accelerators with host DDR and device HBM for large model AI training and inference. MI300X preserves a traditional host and device memory model common in today's machine learning frameworks and applications. Based on the 4th Gen Infinity architecture, MI300X connects eight AMD CDNA™ 3 architecture-based GPU chiplets into a unified AMD Infinity Cache which maintains hardware coherence between the multiple chiplets. The AMD Infinity Fabric™ network on chip (NoC) connects the Infinity Cache to the HBM3 memory and to the external Infinity Fabric and PCIe® Gen 5 links. MI300X is designed for industry-standard open accelerator infrastructure (OAI) UBB2-OAM2 [8,9] systems.

MI300A (Fig. 11.1.7) is targeted at the highest-density HPC systems and combines three AMD "Zen 4" CPU chiplets with six AMD CDNA 3 GPU chiplets to create a heterogeneous computing system in the AMD Infinity Fabric NoC, sharing a unified AMD Infinity Cache backed by HBM3 DRAM. This allows pointers to be passed between CPU and GPU kernels seamlessly without the need for explicit copies or migrations, reducing GPU programming overhead and enabling performance optimizations [4,5]. This level of integration eliminates the need for separate CPU sockets, DDR memory and interconnects between the CPU and GPU, improving power efficiency and is an example of the "product portfolio multiplier" [3] made possible by chiplet technology.

The Instinct MI300 accelerators are the follow-up to the MI250 series and unify 8-stacks of HBM3 as a single high-performance memory system and presents the accelerated compute assets as a single GPU physical device. Methods for cooperative dispatch and synchronization, used to coordinate multiple compute elements, allow the benefits of AMD chiplet architecture [3] to be applied to the GPU for cost advantages, market agility, and reuse. Two new chiplet types are introduced in MI300, the input/output die (IOD) and the accelerator complex die (XCD). The IOD incorporates HBM3 memory controllers, AMD Infinity Fabric and Infinity Cache, IO subsystem, and low power and low latency chip-to-chip ultra-short reach (USR) PHY. The IOD is designed to host 3D hybrid bonded XCD and core complex die (CCD) [7] chiplets. Interconnections between multiple IOD and HBM is accomplished using 2.5D Si interposer-based packaging. The XCD implements the AMD CDNA 3 architecture in a new level of compute GPU hierarchy, the accelerator compute complex (XCC). MI300 uses the same "Zen 4" CCD [1] as the 4th Gen AMD EPYC™ processor [2] with modifications to enable hybrid bonding to the IOD [7]. All "Zen 4" CCD are connected to the IOD with two connections [2] for the highest possible BW to the MI300 Infinity Fabric. MI300A is constructed with four input/output die combined with six XCD chiplets and three CCD chiplets (Fig. 11.1.7). MI300X is constructed with four input/output die combined with eight XCD chiplets.

MI300 leverages XCC hierarchy and synchronization capabilities to introduce spatial partitioning. One or more XCCs are grouped with video/image decode and DMA resources to form partitions. This scheme lends itself to PCIe® single root I/O virtualization (SR-IOV) where each PCIe® virtual function (VF) can be mapped to a partition.

The AMD Infinity Cache is a 256MB memory side cache, designed with an emphasis on power and area efficiency shared among all the clients in the system. Power efficiency is increased significantly due to reduction in DRAM activity, while lower latencies versus HBM, help reduce queueing throughout the data path. Physically, the 256MB Infinity Cache is evenly split between the four IODs and each IOD's Infinity Cache is further divided into 64 1MB tiles, two per HBM channel. The SRAM arrays use the high-density (HD) bitcell for its area and power advantages. The HD bitcells are powered by an on-die regulated VDDM supply. The array macros are co-optimized with the SoC power delivery scheme, sized so the pitched vertical power TSV stripes fit in the channels between the array macros (Fig. 11.1.4).

MI300 is enabled by a grounds-up 3D hybrid bonded [7] architecture (Fig. 11.1.1), unlike the previous instantiation of hybrid bonding in the 3D V-Cache™ product family [6], where the SRAM L3D silicon was stacked to the CCD to provide performance uplift relative to the non-stacked part. MI300 extends the hybrid bonding technology envelope of TSMC's SoIC© platform by enabling multiple top die to be hybrid bonded to approximately half a reticle sized base die. AMD continues to use the face-to-back hybrid bonding architecture with TSV and hybrid bond pitches remaining at 9µm minimum pitch. Unlike the original 3D V-Cache, the bond pad via (BPV), which connects the bond pad metal (BPM) to the metal stack on the top die, now lands on an aluminum redistribution layer instead of the top-most copper layer to improve power delivery (Fig. 11.1.2). Each IOD contains over 350,000 power and signal TSVs, totaling over 1.4 million hybrid bonded connections per package. The extreme density TSV integration is enabled by the very fine hybrid bonding pitch and low overhead that hybrid bonding supports. MI300 is not just the first AMD multiple die hybrid bonded architecture but also the first AMD hybrid bonded 3D + 2.5D architecture, utilizing the Silicon interposer to make fine micro-bump connections between the IOD-IOD and IOD-HBM.

The advanced packaging and chiplet architecture allow each element of MI300 to be designed in the most appropriate technology, optimizing for power, performance, and cost. The MI300 IOD is manufactured in TSMC N6 and contains 12.2 billion transistors and the XCD is manufactured in TSMC N5 and contains 13 billion transistors. MI300 IOD is a two GDS design, where one of the IOD is a minimally modified mirrored copy of the original. USR TX and RX ordering is swapped on the mirrored IOD to align the TX to RX of the original IOD and hybrid bonding sites for signals in the IOD must account for mirrored IODs with non- mirrored top dies as shown in Fig. 11.1.3.

High-speed IOD-to-IOD interconnect is achieved via the USR PHY using several metal layers of the Silicon interposer. The USR interconnect has a minimum micro-bump pitch of 35µm, while the IOD-to-HBM interconnect has a minimum pitch of 45µm. Miscellaneous IOD-to-IOD interconnects are routed on the substrate which allow the AMD Instinct MI300A and MI300X accelerators to share the same silicon interposer design. The USR PHY enables greater than 10× higher areal bandwidth density compared to traditional SerDes [3] while reducing power consumed by these short (~1 mm) cross-die links. The PHY is composed of modular TX/RX and CLK blocks running at 8.4 Gbps (Fig. 11.1.3) with optimized circuits to fit within the dense 35µm micro-bump pitch and using unterminated CMOS signaling without equalization or termination to allow direct power scaling with bandwidth. Fig. 11.1.5 shows a representative bump pattern with the TX and RX phylets highlighted. The design uses a clock-forwarded die-die interface with two-phase Inline/quadrature clocks from a central PLL. Precisely matched data and clock delays enable high tolerance of cross-die PVT and device voltage variations. Per-lane de-skew is implemented to account for remaining variation. The PHY is designed to operate along a noisy supply environment, with tolerance enabled by the carefully matched delays and simple CMOS circuits, without equalization, termination, or CDR. USR power is significantly reduced compared to traditional SerDes, including for workloads with partial bandwidth enabled by aggressive fine-grain clock gating. Clock-gated power is 10% of active 50% toggling power and there is no RX power for cycles when gated. The USR design achieves measured power of 0.4mW/Gbps with 4.38Tbps/mm^2 areal bandwidth.

Power delivery to the top dies utilizes a uniform TSV pattern in the IOD (Fig. 11.1.4) that can deliver greater than 1.5A/mm^2 with minimal I^{2}R power loss. In addition to delivering power to the top dies through the TSVs, the IOD micro-bump pattern at the silicon interposer interface can simultaneously deliver more than 0.5A/mm^2 to the IOD logic. As the device transitions between compute intensive and memory intensive workloads, power can be shifted between the top dies and the IOD.

All four IOD chiplets share the same reference clock which is distributed on the substrate. The MI300 PCIe® subsystem can act as either a root complex or an end-point and thus, the reference clock scheme is configurable depending self-hosted or end-point device. The primary IOD configures the appropriate reference clock scheme and provides the reference clock to all other secondary IODs.

The AMD Instinct MI300X accelerator's advanced packaging, process nodes, and chiplet technologies combine to enable a single GPU accelerator device with a unified 8-stack HBM3 memory system. This delivers over 2.5× matrix FMA FP16 KOPS/CLK and 1.5× HBM capacity and pin width*rate (Fig. 11.1.6) versus the aggregate of the two GPU compute dies (GCD) of the AMD MI250X (Fig. 11.1.9) [10]. The MI300 modular chiplet package enables the MI300A self-hosted APU and the MI300X PCIe® OAM to power next-generation AMD HPC and AI platforms.

References:

[1] B. Munger, et al., "Zen 4": The AMD 5nm 5.7GHz x86-64 Microprocessor Core," *ISSCC*, pp. 38-39, Feb. 2023.

[2] 4TH GEN AMD EPYC™ PROCESSOR ARCHITECTURE, Available online: <https://www.amd.com/system/files/documents/4th-gen-epyc-processor-architecture-white-paper.pdf>, Sept. 2023.

[3] S. Naffziger, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families: Industrial Product," *ACM/IEEE Int. Symp. Computer Architecture*, pp. 57-70, 2021.

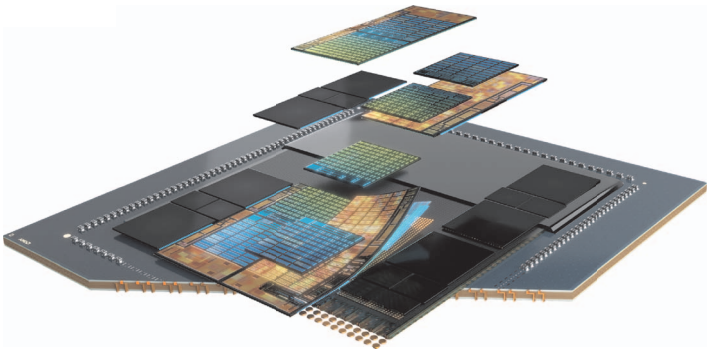

Figure 11.1.1: MI300X 2.5D and 3D Packaging.

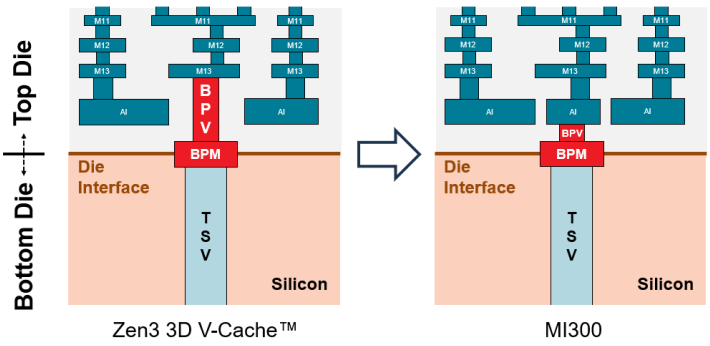

Figure 11.1.2: MI300 vs. "Zen 3" 3D V-Caches.

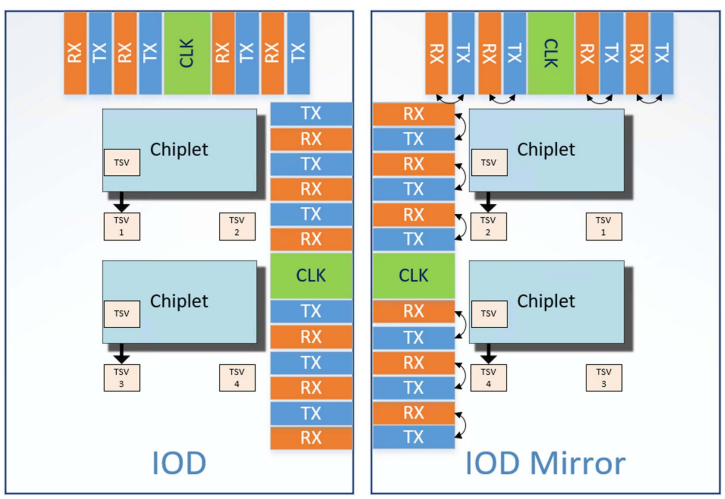

Figure 11.1.3: IOD Mirroring and USR PHY Blocks.

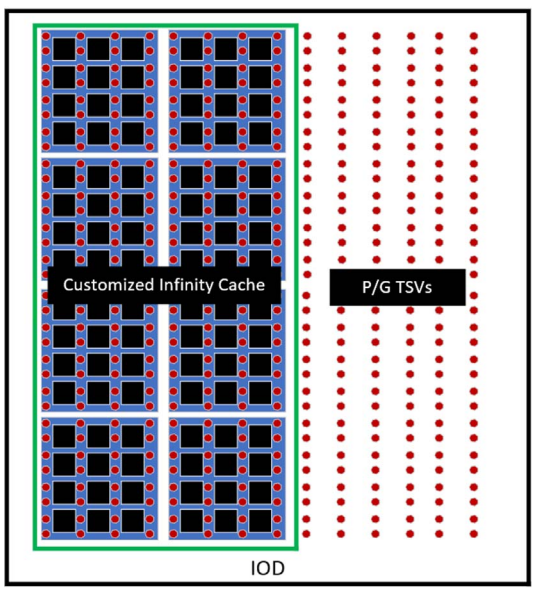

Figure 11.1.4: IOD Uniform TSV Pattern.

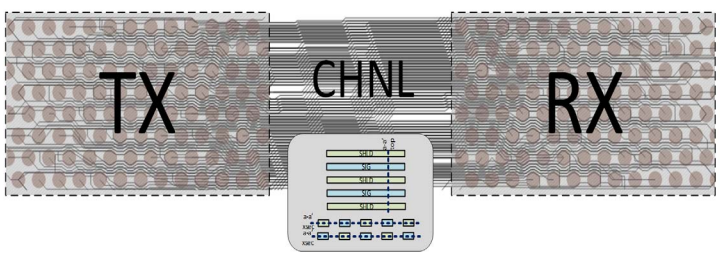

Figure 11.1.5: USR Bump Pattern and Fanout Routes.

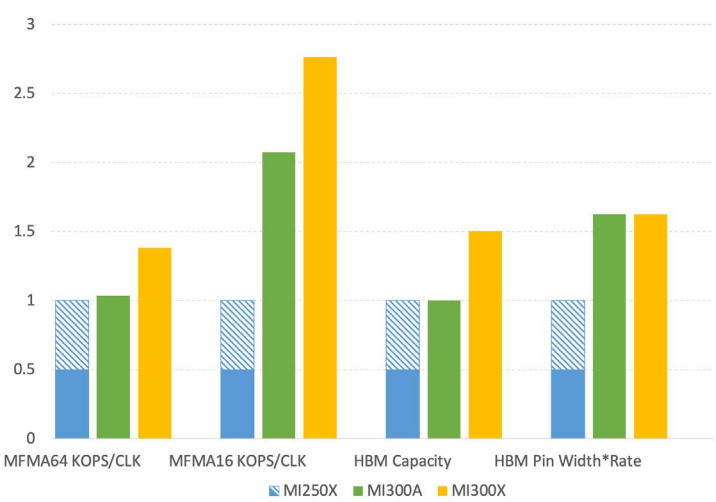

Figure 11.1.6: MI300 Throughput and Capacity vs. MI250X.

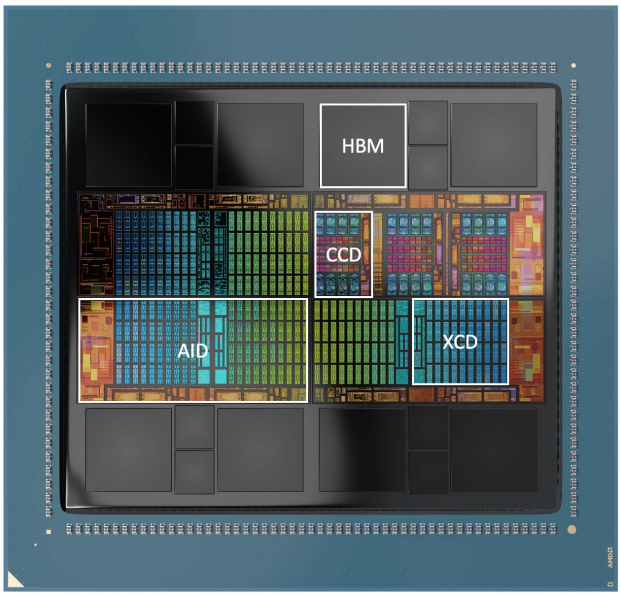

Figure 11.1.7: MI300A Module and Chiplets Photograph.

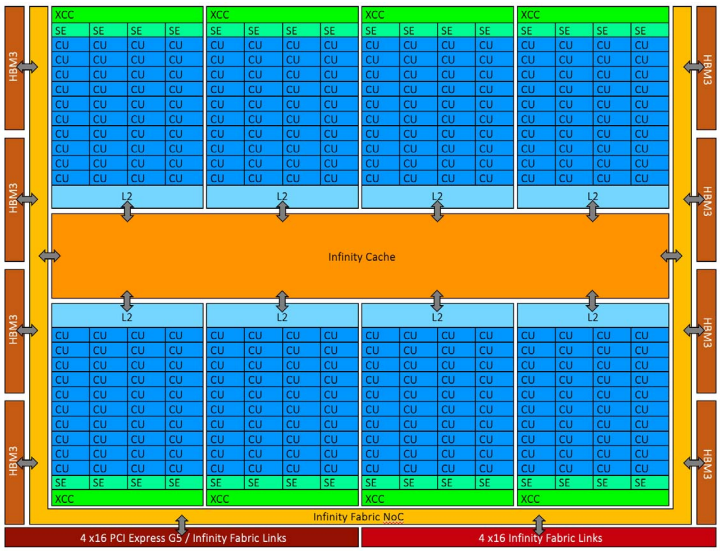

Figure 11.1.8: MI300X Block Diagram.

Product	AMD Instinct™ MI250X	AMD Instinct™ MI300X	NVIDIA H100 SXM
GPU Architecture	AMD CDNA™ 2	AMD CDNA™ 3	NVIDIA Hopper
Lithography	TSMC 6nm FinFET	TSMC 6nm, TSMC 5nm	TSMC 4N
Total Board Power (TBP)	670W	750W	700W
Peak Engine Clock	1700 MHz	2100 MHz	1980 MHz FP64 1830 MHz BF16
Peak DP (FP64) Performance	47.9 TFLOPS	81.7 TFLOPS	33.5 TFLOPS
Peak DP Matrix (FP64) Performance	95.7 TFLOPS	163.4 TFLOPS	66.9 TFLOPS
Peak bfloat16 Matrix Performance	383 TFLOPS	1307 FLOPS	989.4 TFLOPS
Memory Type	HBM2e	HBM3	HBM3
Memory Clock	1.6 GHz	2.6 GHz	2.619 GHz
Memory Interface	8192-bit	8192-bit	5120-bit
Peak Memory Bandwidth	3276.8 GB/sec	5324.8 GB/sec	3352 GB/sec

Figure 11.1.9: Comparison Table [10, 11].

Additional References:

[4] T. Vijayaraghavan, et al., "Design and Analysis of an APU for Exascale Computing," *Int. Symp. High Performance Computer Architecture*, pp. 85-96, 2017.

[5] M.J. Schulte, et al., "Achieving Exascale Capabilities Through Heterogeneous Computing," *IEEE Micro*, vol. 35, no. 4, pp. 26-36, July-Aug. 2015.

[6] J. Wuu, et al., "3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU," *ISSCC*, pp. 428-429, Feb. 2022.

[7] R. Agarwal, et al., "3D Packaging for Heterogeneous Integration," *IEEE Electronic Components and Technology Conf.*, pp. 1103-1107, 2022.

[8] Open Accelerator Infrastructure (OAI) - Universal Baseboard (UBB), Available online: <https://www.opencompute.org/documents/oai-ubb-base-specification-r2-0-v0-5-2-pdf>, Mar. 2023.

[9] Open Accelerator Infrastructure (OAI) - OCP Accelerator Module (OAM), Available online: <https://www.opencompute.org/documents/oai-oam-base-specification-r2-0-v0-75-2-pdf>, Mar. 2023.

[10] A. Smith and N. James, "AMD Instinct™ MI200 Series Accelerator and Node Architectures," *IEEE Hot Chips 34*, pp. 1-23, 2022.

[11] NVIDIA H100 Tensor Core GPU Architecture, Available online: <https://resources.nvidia.com/en-us-tensor-core>, 2023.

979-8-3503-0620-0/24/$31.00 ©2024 IEEE