IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF TEXAS
MIDLAND-ODESSA DIVISION

FERMAT INTERNATIONAL, INC., Plaintiff, v. ADVANCED MICRO DEVICES, INC., Defendant.	Civil Action No. 7:26-cv-00011 **JURY TRIAL DEMANDED**

COMPLAINT FOR PATENT INFRINGEMENT

Plaintiff Fermat International, Inc. ("Fermat") files this Complaint against Advanced Micro Devices, Inc. ("AMD") for patent infringement of United States Patent Nos.: 11,550,512; 11,886,750; and 12,271,636 (collectively the "Patents-in-Suit" and attached hereto as Exhibits 1-3 respectively), and alleges as follows:

NATURE OF THE ACTION

1. This is an action for patent infringement arising under the patent laws of the United States, 35 U.S.C. §§ 1 *et seq.* and seeking damages and injunctive relief as provided in 35 U.S.C. §§ 281 and 283-285.

THE PARTIES

2. Fermat is a corporation organized under laws of the State of Delaware with its principal place of business at 2625 Middlefield Road #800, Palo Alto, California 94306.

3. Fermat was founded by inventors Robert Bismuth and Mike Stengle in 2017. The inventors sought to solve problems with "Big Data," high performance computing, and solid-state device technologies.

4. At the time of the inventions of the Patents-in-Suit, the rate at which new data

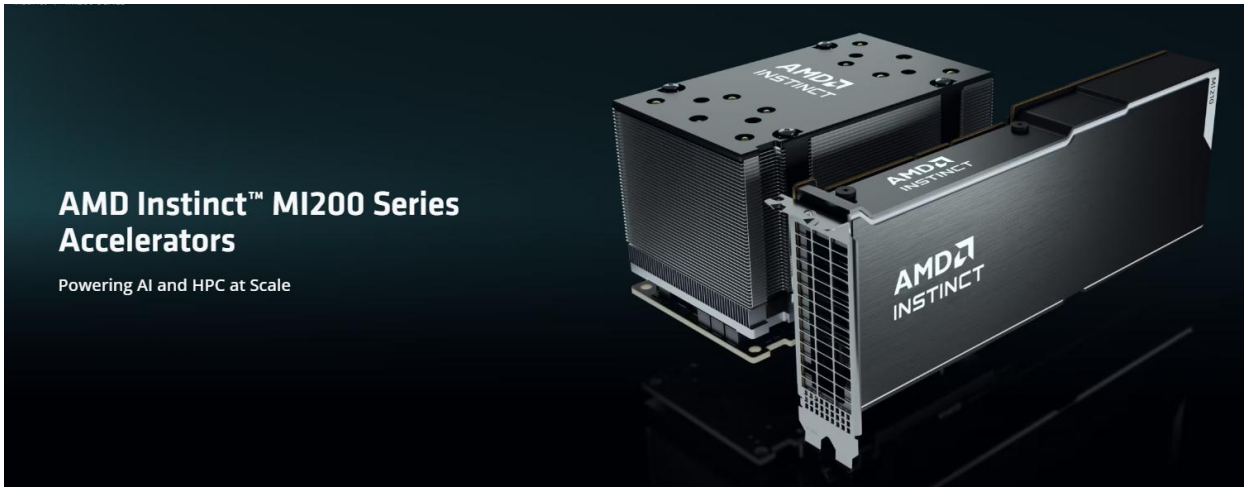
AMD 1062
Advanced Micro Devices v. Fermat Int'l
IPR2026-00358

became available for processing outpaced the rate at which processing methodologies of traditional hardware and software systems were advancing to process that data. Thus, it would take longer to solve bigger and more complex data processing problems.

5. Fermat's patented technology enables high-performance computation for big data analytics. Fermat provides a "DATA AWARE" High-Performance Computing Solution that scales linearly with data volume in order to solve the issues created when processing vast amounts of data.

6. On information and belief, Defendant Advanced Micro Devices, Inc. is a corporation organized under the laws of the State of Delaware, with its principal place of business at 2485 Augustine Drive, Santa Clara, California 95054. AMD may be served with process through its registered agent, CT Corporation System, 1999 Bryan St., Suite 900, Dallas, Texas 75201.

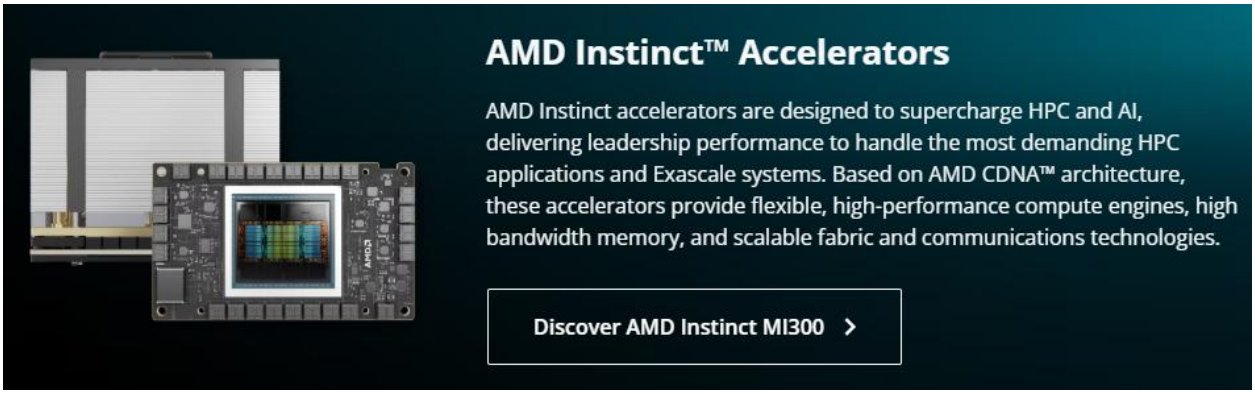
7. AMD's Instinct™ Accelerators (the "Accused Products"), such as the MI200 series and the MI300 series, are data processing systems that enable exceptional performance for a data center, at scale.



Source: AMD Instinct™ MI200 Series Accelerators.

8. AMD's Instinct™ Accelerators are designed to supercharge high-performance computing and AI, and are based on AMD's CDNA™ architecture. The accelerators provide

flexible, high-performance compute engines, high bandwidth memory, and scalable fabric and communications technologies.



Source: HPC Solutions.

JURISDICTION AND VENUE

9. Fermat asserts claims for patent infringement arising under the patent laws of the United States, 35 U.S.C. § 100, *et seq.* This Court has subject matter jurisdiction of these claims under, *inter alia*, 28 U.S.C. §§ 1331 and 1338(a).

10. This Court has personal jurisdiction over AMD in this action, because AMD has established minimum contacts with Texas as a whole, such that the exercise of jurisdiction would not offend traditional notions of fair play and substantial justice. AMD has purposefully directed activities at Texas and engages in sales and marketing efforts to generate and support such sales. On information and belief, AMD directly and through subsidiaries, intermediaries, and third parties, has committed and continues to commit acts of infringement in this District by, among other things, making, using, offering to sell, selling, and/or importing products that infringe the Patents-in-Suit.

11. Venue is proper in this District under 28 U.S.C. §§ 1391 and 1400(b). Upon information and belief, AMD has transacted business in this District and has committed acts of direct and indirect infringement in this District by, among other things, making, using, offering to

sell, selling, and/or importing products that infringe the Patents-in-Suit. AMD has regular and established places of business in this District, including offices at 1340 Airport Commerce Drive, Austin, Texas 78741; 7000 West William Cannon Drive Austin, Texas 78735; 7171 Southwest Parkway Austin, Texas 78735; and 12301 Research Blvd., Suite 300 Austin, Texas 78759.

12. On information and belief, as of July 2025, AMD had more than 3,700 employees in this District. AMD has extensive facilities with documents and thousands of employees in this District, and maintains a permanent and/or continuing presence within this District. On information and belief, several current and former AMD employees with knowledge relevant to the claims in this case are located in this District and within the State of Texas. For example, current AMD employees whose job titles or job experience is related to the AMD Instinct™ MI200 Series Accelerators are located in this District.

13. AMD's Instinct™ MI200 Series Accelerators have previously been accused of infringement in this District and several current and former AMD employees with knowledge relevant to the claims in this case were identified as located in this District. AMD did not contest venue in this District and in this division. *See Advanced Cluster Systems, Inc. v. Advanced Micro Devices, Inc*., No. 7:24-cv-00244 (W.D. Tex. Sept. 26, 2024) (Dkts. 1, 29).

THE PATENTS-IN-SUIT

14. The Patents-in-Suit are United States Patent Nos. 11,550,512 ("the '512 patent"); 11,886,750 ("the '750 patent"); and 12,271,636 ("the '636 patent").

15. Each of the Patents-in-Suit is presumed valid under 35 U.S.C. § 282.

16. Fermat owns all rights, title, and interest in each of the Patents-in-Suit, including the right to recover for all past and ongoing infringement.

17. Fermat has not granted AMD an approval, an authorization, or a license to any

rights under any of the Patents-in-Suit.

18. Fermat (and any predecessors and/or licensees) complied with 35 U.S.C. § 287, because Fermat (or any predecessor/and or licensee) did not make, offer for sale, sell, or import into the United States any products that practice the Patents-in-Suit during the relevant time period or was not otherwise required to mark.

19. The claimed inventions in the Patents-in-Suit sought to solve problems with, and improve upon, high performance data processing and, more particularly, distributed hardware architecture. To enable high performance data processing, a distributed hardware architecture is provided to manage and execute data processing operations. One or more compute nodes in the data processing system are utilized to manage and execute the data processing operations independent of, though communicatively coupled with, a host compute system.

United States Patent No. 11,550,512

20. The ’512 patent, entitled “Analytics, algorithm architecture, and data processing system and method,” was duly and legally issued by the United States Patent and Trademark Office (“USPTO”) on January 10, 2023 to inventors Robert Bismuth and Mike Stengle.

21. The ’512 patent is directed to a specific technological improvement in high performance data processing. The invention presents a concrete and specific solution to technical problems associated with conventional systems. In particular, the invention pertains to an improved system and method employing a distributed hardware architecture that facilitates high throughput data processing and analytics solutions for resource-intensive applications.

22. The claims of the ’512 patent recite a specific and inventive combination of elements that improve data processing in a manner that is not routine or conventional. Unlike standard techniques, the claimed invention of the ’512 patent recites a distributed hardware

architecture wherein one or more compute nodes in the data processing system are utilized to manage and execute the data processing operations independent of, though communicatively coupled with, a host compute system.

23. The claims recite specific, particularized, unconventional, technical solutions, including, for example: a plurality of serially coupled compute nodes in an execution pipeline; a communications link, wherein a respective compute node is communicatively coupled to the node interface at the router module via the communications link either directly or indirectly via an adjacent compute node; a data store comprising records associated with a data processing operation; a programmable logic component to execute the data processing operation in cooperation with the data store; a node memory comprising data and instructions to support operation of the programmable logic component in connection with the data processing operation and to facilitate inter-node data communications via the communications link; a data mover component to facilitate intra-node data communications between the programmable logic component and the node memory at the respective compute node; and a storage interface component to facilitate intra-node data communications between the programmable logic component and the data store at the respective compute node.

24. The distributed hardware architecture of the ’512 patent claims presents specific, non-conventional and non-routine technical advancements for processing large datasets in a way that enhances efficiency and performance. The claimed invention departs from traditional approaches that use overwhelming computational resources to process large datasets, at the expense of cycle time and power consumption.

United States Patent No. 11,886,750

25. The ’750 patent, entitled “Analytics, algorithm architecture, and data processing

system and method," was duly and legally issued by the USPTO on January 10, 2023 to inventors Robert Bismuth and Mike Stengle.

26. Like the '512 patent, the claims of the '750 patent address specific, non-conventional and non-routine technical advancements for data processing large datasets in a way that enhances efficiency and performance.

United States Patent No. 12,271,636

27. The '636 patent, entitled "Analytics, algorithm architecture, and data processing system and method," was duly and legally issued by the USPTO on April 8, 2025 to inventors Robert Bismuth and Mike Stengle.

28. Like the '512 patent and '750 patent, the claims of the '636 patent address specific, non-conventional and non-routine technical advancements for data processing large datasets in a way that enhances efficiency and performance.

CLAIMS FOR RELIEF

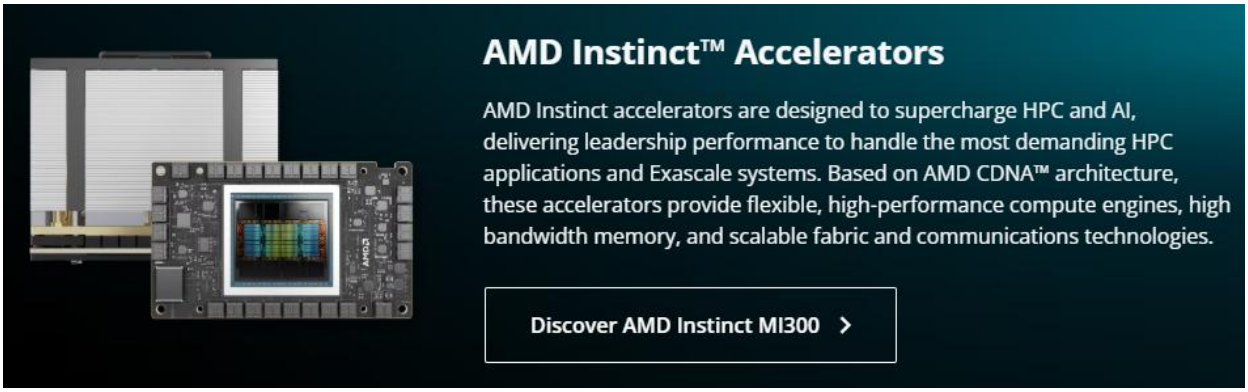
Count I – Infringement of United States Patent No. 11,550,512

29. Fermat repeats, realleges, and incorporates by reference, as if fully set forth here, the allegations of the preceding paragraphs above.

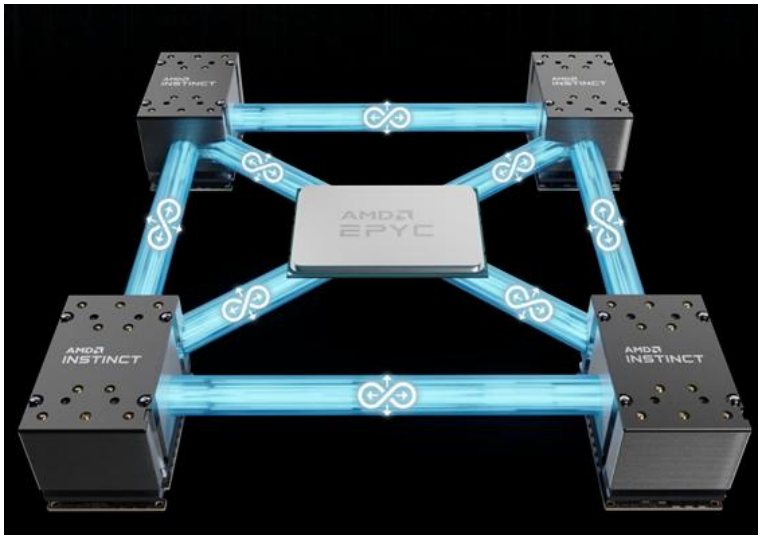
30. AMD has directly infringed and continues to infringe, literally and/or through the doctrine of equivalents, at least claim 1 of the '512 patent by making, using, selling, offering for sale, and/or importing one or more of the Accused Products, in violation of 35 U.S.C. § 271(a).

31. To the extent the preamble is limiting, the AMD Instinct™ MI200 Series Accelerator is a data processing system operative in cooperation with a host compute system. The AMD Instinct™ MI200 Series Accelerator is used for, *e.g.*, High Performance Computing (HPC) and Artificial Intelligence (AI) based applications, which involves various data processing

operations. The AMD Instinct™ MI200 Series Accelerator operates in cooperation with the Central Processing Unit ("CPU") ("host compute system"), such as AMD's EPYC.



Source: HPC Solutions.



Source: AMD Instinct™ MI200 Series Accelerator and Node Architectures.

32. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a router module, which includes at least an Infinity Fabric Link that provides the connection to a CPU (*e.g.*, EPYC) and a controller that controls the routing of data over the Infinity Fabric. The Infinity Fabric Link (which is part of a router module) comprises a host interface and a node interface, wherein the router module is communicatively coupled to the host compute system via the host interface.

33. The AMD Instinct™ MI200 Series Accelerator can be coupled to (and is in communications with) a CPU (*e.g.*, EPYC) through the host interface of the Infinity Fabric Link.

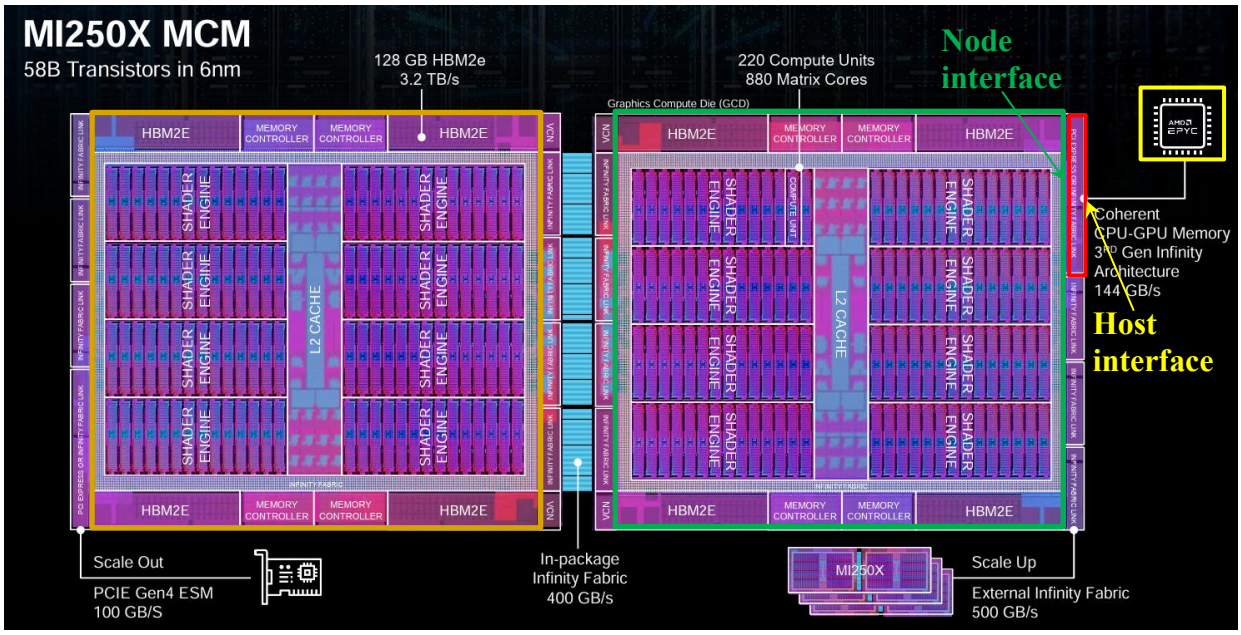
> Key capabilities and features of the AMD Instinct MI200 series accelerators include:
>
> - **AMD CDNA™ 2 architecture** – 2nd Gen Matrix Cores accelerating FP64 and FP32 matrix operations, delivering up to 4X the peak theoretical FP64 performance vs. AMD previous gen GPUs. [1,3,4]
> - **Leadership Packaging Technology** – Industry-first multi-die GPU design with 2.5D Elevated Fanout Bridge (EFB) technology delivers 1.8X more cores and 2.7X higher memory bandwidth vs. AMD previous gen GPUs, offering the industry's best aggregate peak theoretical memory bandwidth at 3.2 terabytes per second. [4,5,6]
> - **3rd Gen AMD Infinity Fabric™ technology –** Up to 8 Infinity Fabric links connect the AMD Instinct MI200 with 3rd Gen EPYC CPUs and other GPUs in the node to enable unified CPU/GPU memory coherency and maximize system throughput, allowing for an easier on-ramp for CPU codes to tap the power of accelerators.

Source: New AMD Instinct™ MI200 Series Accelerators Bring Leadership HPC and AI Performance to Power Exascale Systems and More.

34. The Infinity Fabric Link is shown in the red box below. As shown below, the Infinity Fabric Link comprises a host interface and a node interface. The host interface connects to the host compute system (*e.g.*, EPYC) (shown in the yellow box below) and the node interface is communicatively coupled to the Graphic Compute Dies ("GCDs") (shown in green and orange boxes below) ("compute node") of the AMD Instinct™ MI200 Series Accelerator.

> The coherent host interface is another novel aspect of the AMD CDNA 2 architecture, which enables the advanced memory coherency. The physical layer is implemented as a 16-lane Infinity Fabric link. Logically the link can behave like an Infinity Fabric interface when paired with an optimized 3rd Gen AMD EPYC™ processor, enabling the unique cache coherency capabilities. When connected with other X86 server processors, this interface falls back to behave like a standard PCIe interface with non-coherent communication to the host processor.

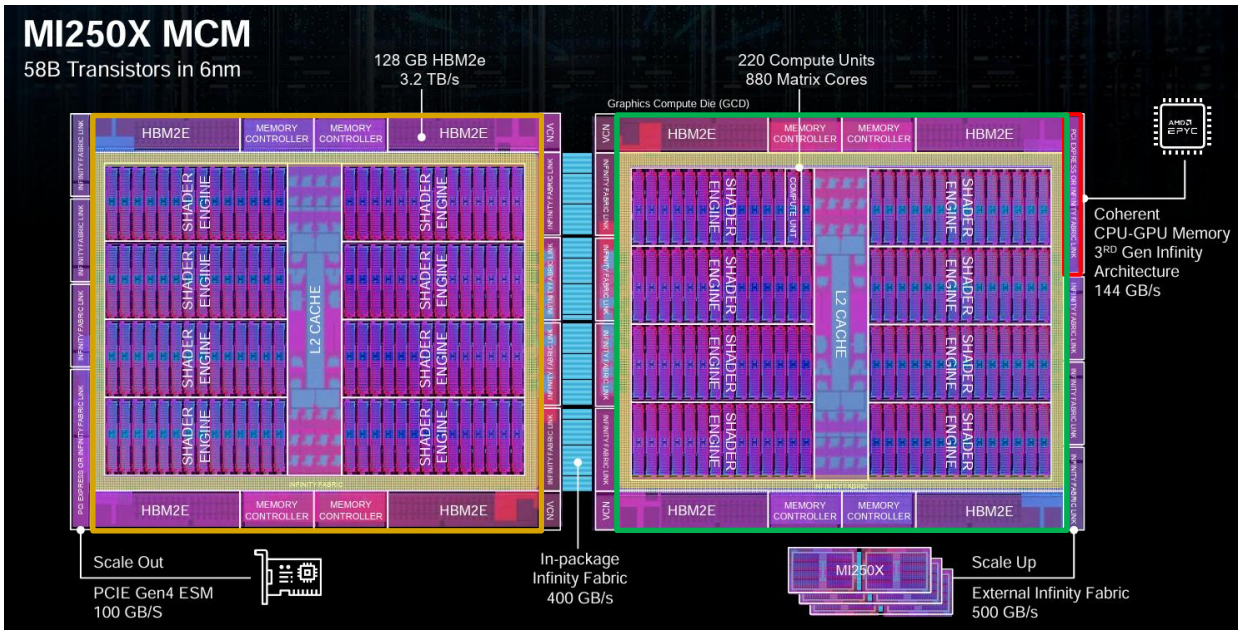
Source: AMD CDNA2 White Paper.



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

35. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline. The AMD Instinct™ MI200 Series Accelerator is a multi-die Graphic Processing Unit ("GPU") as shown above. Pairs of GCDs ("compute nodes") in each AMD Instinct™ MI200 Series Accelerator are serially coupled in an execution pipeline. For example, as shown above, the left GCD (shown in the orange box) is connected to the right GCD (shown in the green box) using the In-package Infinity Fabric. The right GCD, in turn, is connected to the EPYC.

36. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: a communications link, wherein the respective compute node is communicatively coupled to the node interface at the router module via the communications link either directly or indirectly via an adjacent compute node.



Source: AMD Instinct™ MI200 Series Accelerator and Node Architectures.

37. As shown above, the internal Infinity Fabric (yellow shaded area) is/includes a communications link(s). The communications link can also include the In-package Infinity Fabric. The compute node (GCD, green box) is communicatively coupled directly to the node interface of the Infinity Fabric Link (red box) (which is part of the router module) via a communications link. Moreover, the compute node (GCD, orange box) is communicatively coupled to the node interface via a communications link indirectly through the adjacent compute node (GCD, green box).

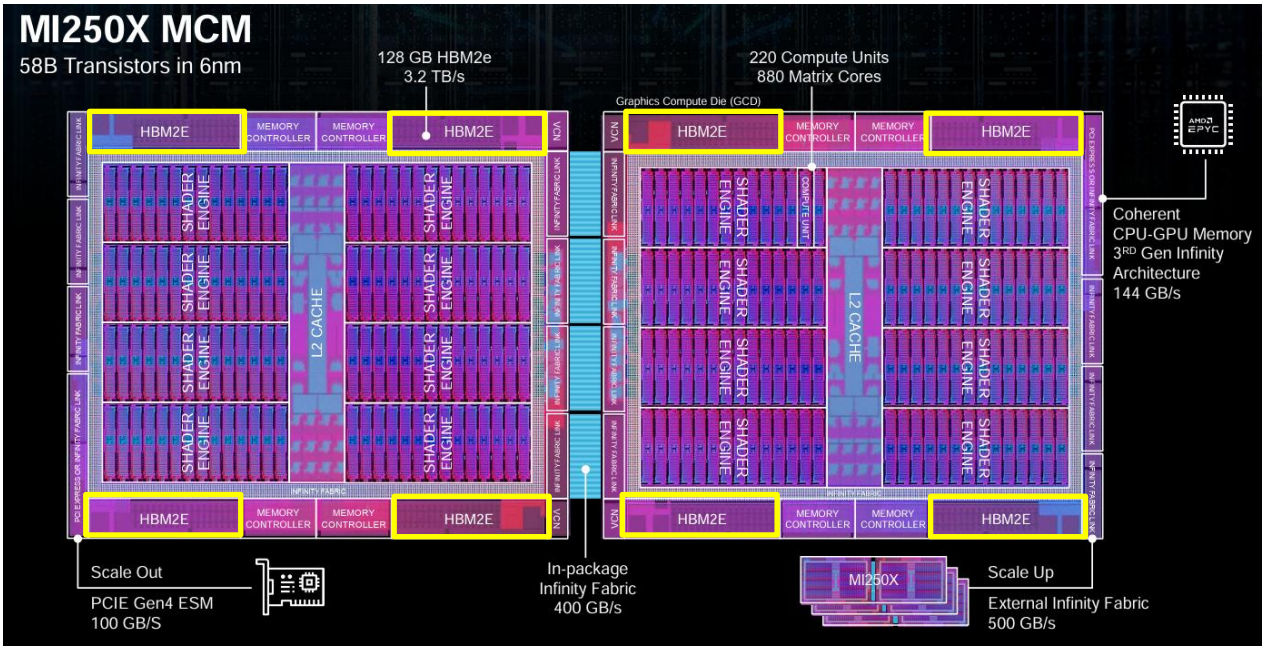
Communication and Scaling

In many ways, the most critical improvements to the AMD CDNA™ 2 architecture are in the communication capabilities of each GCD within the AMD MI200 device and especially in the unique capabilities offered by AMD Infinity Fabric™ technology. The previous generation relied on standard PCI-Express to connect to the host processor and offered three AMD Infinity Fabric™ links connecting to other GPUs. In the flagship HPC topology example shown in Figure 2a with MI250X, the AMD CDNA 2 architecture builds out the communication capabilities to an entirely different level with four different types of interfaces specialized for different purposes: in-package Infinity Fabric, inter-package Infinity Fabric links, coherent Infinity Fabric links to the host processor, and a downstream PCIe link. The in-package Infinity Fabric, coherent Infinity Fabric, and downstream PCIe links are all novel and unlock the unique system capabilities illustrated in Figure 2a. In the more traditional mainstream and flagship machine learning topologies leveraging the MI250, illustrated in Figure 2b-c, the GPUs are connected to the host processor via PCIe but still benefit from the increased number of GCD-to-GCD Infinity Fabric Links within the GPU device as well as the inter-package external Infinity Fabric links.

Source: AMD CDNA™ 2 Architecture.

38. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective

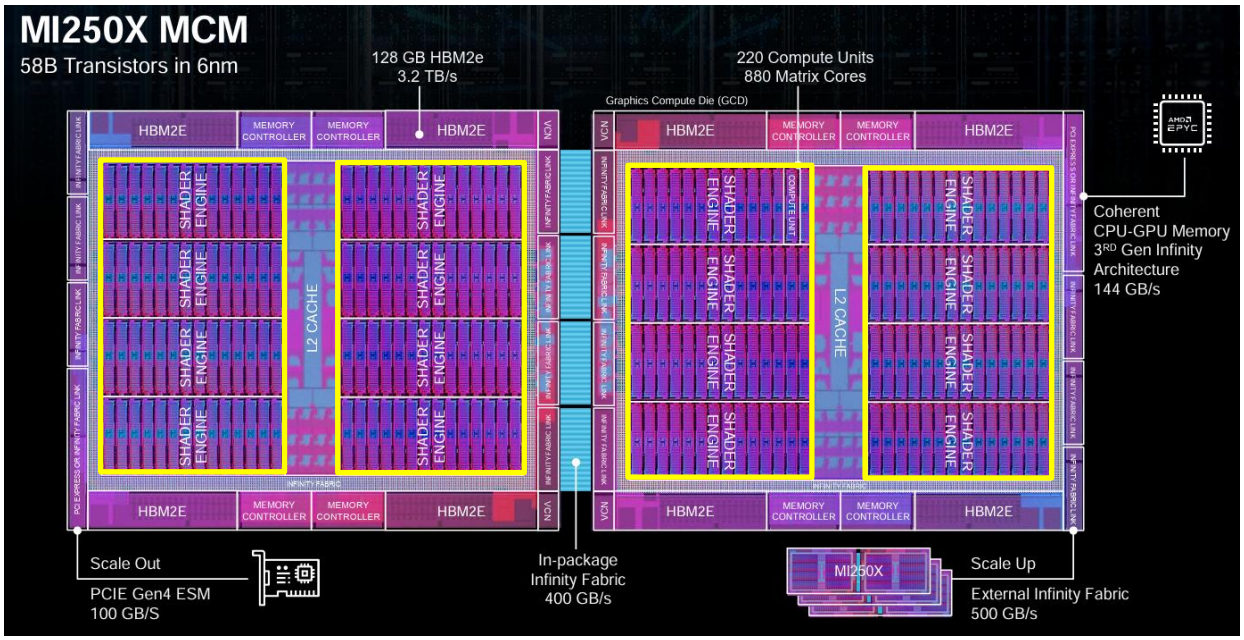
compute node in the execution pipeline comprising: a data store comprising records associated with a data processing operation. The AMD Instinct™ MI200 Series Accelerator employs a multi-level memory hierarchy designed for high-speed data access. High Bandwidth Memory (HBM2e) ("data store") (shown in yellow boxes below) is the main memory for the GCD, where large datasets ("records") and application code reside for a data processing operation (*e.g.*, complex data computations (HPC and AI workloads)). The AMD Instinct™ MI200 Series Accelerator has two GCDs in a single package and each GCD has its own 64 GB of HBM2e memory.



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

39. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: a programmable logic component to execute the data processing operation in cooperation with the data store. As shown below, each GCD comprises Compute Units ("CU")/engines ("programmable logic component") (shown in yellow boxes) that receive data from the HBM2e memory to execute the data processing operation.

The AMD CDNA 2 architecture incorporates 112 physical compute units per GCD, divided into four arrays; the initial products include 104 (for AMD Instinct™ MI250 and the MI210) or 110 (for AMD Instinct™ MI250X) active CUs per GCD. As illustrated in Figure 3, each CU contains register files and pipelines optimized for scalar, vector, and matrix instructions. The vector instructions work with wavefronts that contain 64 work-items and most double precision instructions are executed over four cycles using one set of four pipelines that are each 16-wide. The matrix instructions pull data from the vector register file but have their own specialized data paths that take advantage of implicit data re-use in matrix multiplication to reduce the number of register access for a given computation to help improve both performance and efficiency. Data can be shared between lanes within a wavefront using permute instructions or the local data store (LDS).



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

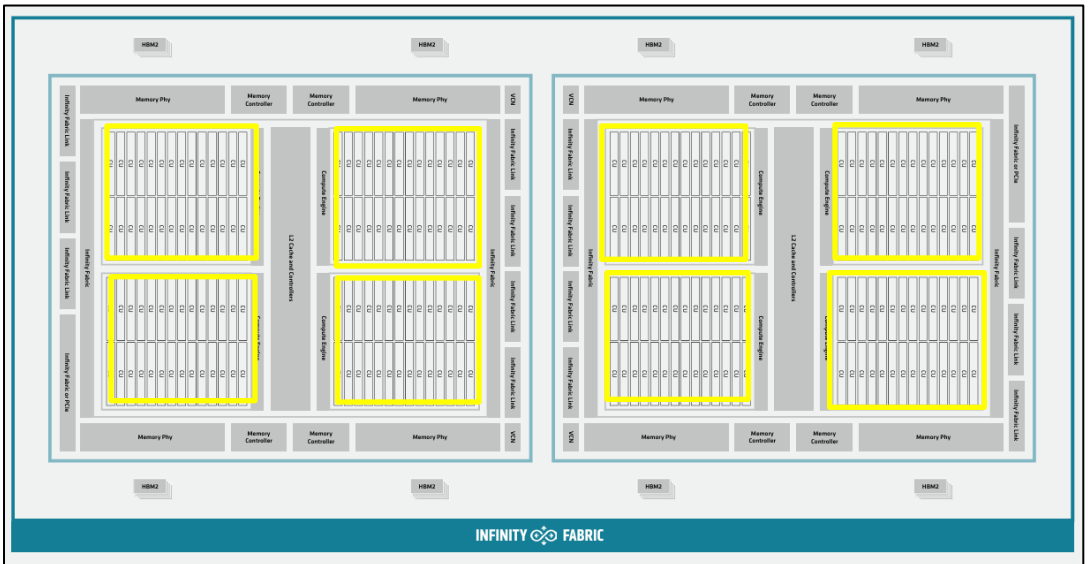
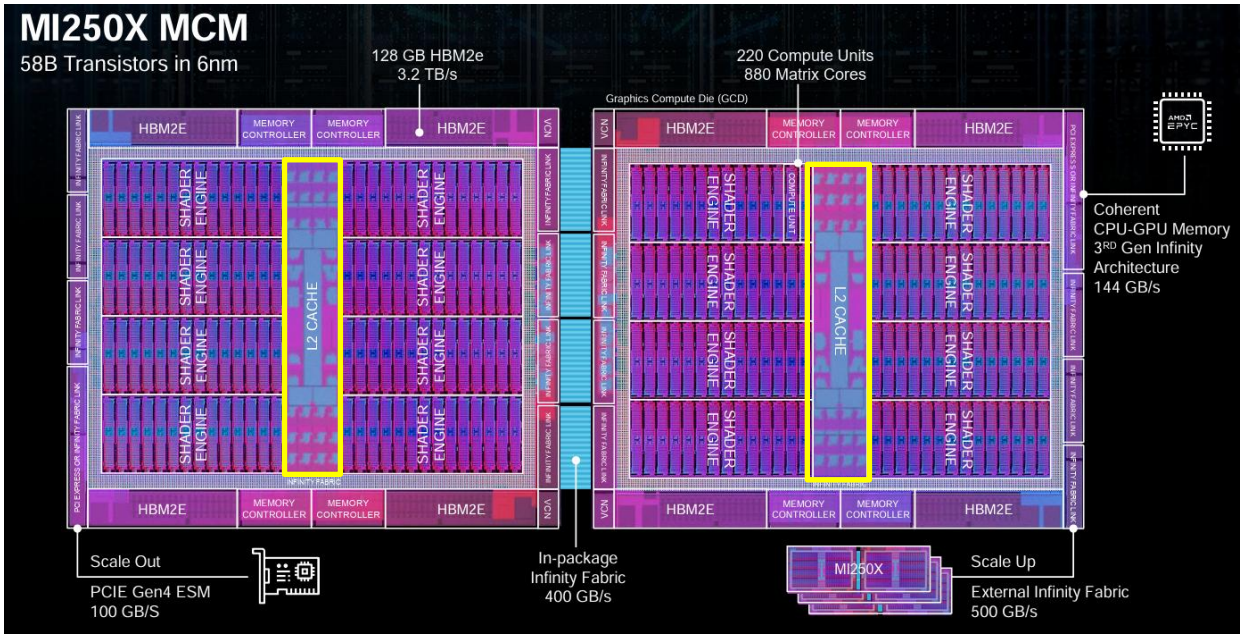


Figure 1b –Block diagram of the AMD Instinct™ MI200 multi-chip module (AMD Instinct™ MI250/MI250X) in the OAM form factor accelerator which comprise two Graphics Compute Dies (GCD) as illustrated.

Source: [AMD CDNA™ 2 Architecture](#).

40. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: a node memory comprising data and instructions to support operation of the programmable logic component in connection with the data processing operation and to facilitate inter-node data communications via the communications link. As discussed above, each GCD has memory resources with data and instructions to support operation of the compute units/engine. The AMD Instinct™ MI200 Series Accelerator employs a multi-level memory hierarchy designed for high-speed data access—L1 cache, L2 cache, and HBM2e. HBM2e is the main memory for the GCD, where large datasets and application code/instructions reside.



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

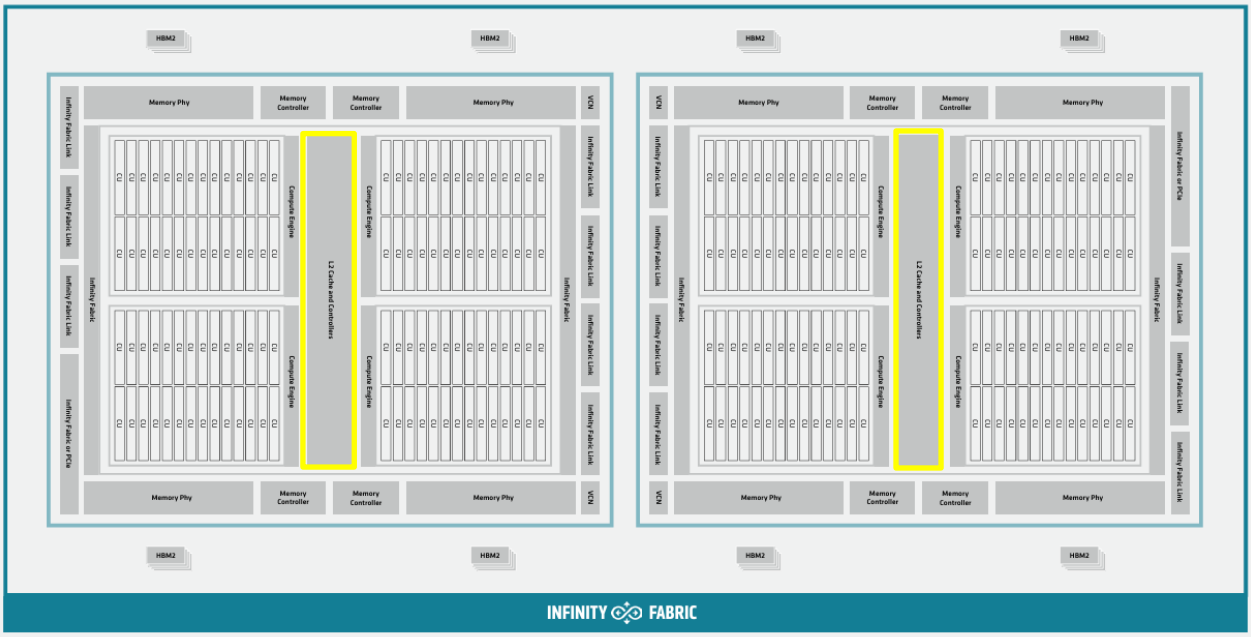


Figure 1b –Block diagram of the AMD Instinct™ MI200 multi-chip module (AMD Instinct™ MI250/MI250X) in the OAM form factor accelerator which comprise two Graphics Compute Dies (GCD) as illustrated.

Source: AMD CDNA™ 2 Architecture.

41. Each GCD has an L2 cache/controller ("node memory") (shown in yellow boxes above) that receives/contains data/instructions from the HBM2e to support operation of the programmable logic component in connection with the data processing operation and to facilitate inter-node data communications via the communications link.

42. In particular, the L2 cache/controller acts as a buffer between the HBM2e and the L1 caches. The L1 caches receives data/instructions from the L2 cache/controller which the programmable logic component uses to perform data processing operations. The L1 cache is the fastest cache in the hierarchy and is located within each CU.

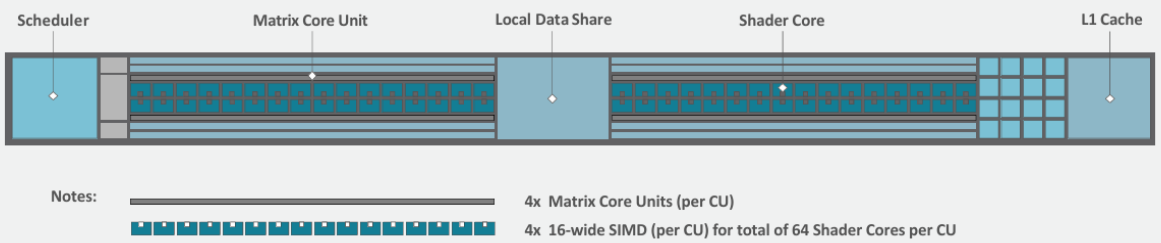
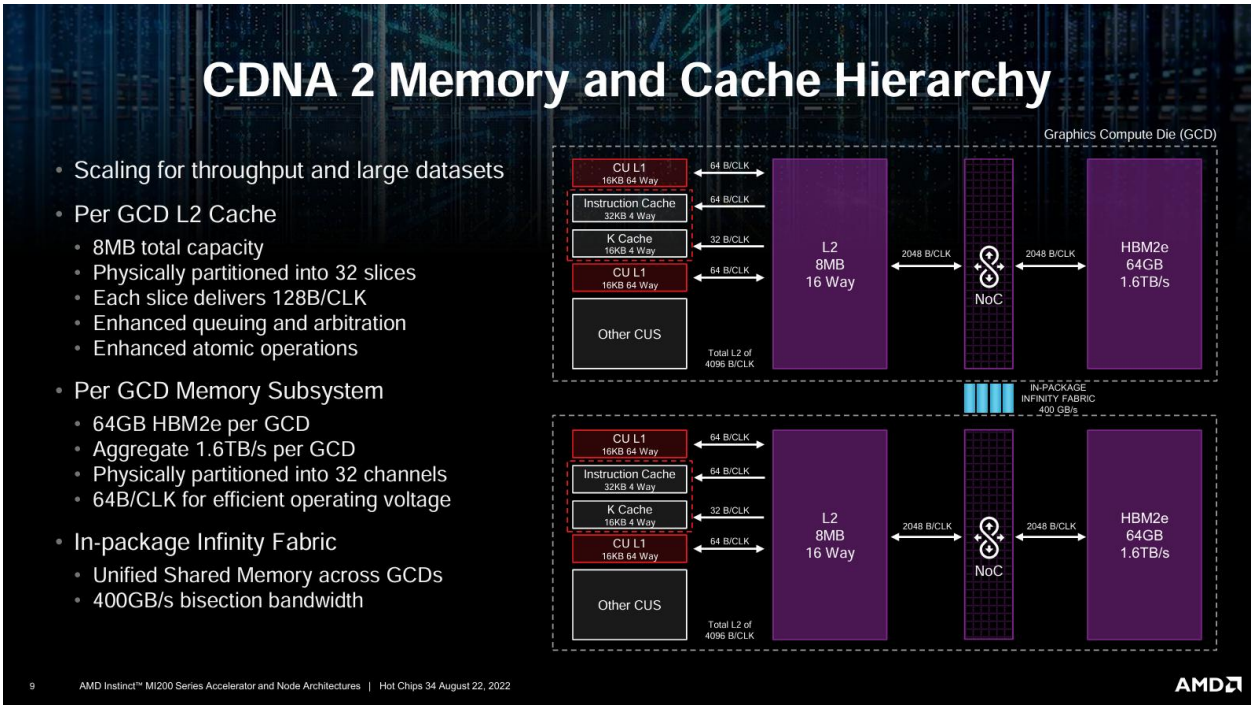


Figure 3 – Conceptional Block Diagram of an Enhanced Compute Unit (CU) with SIMD view of the AMD CDNA™ 2 architecture

Source: AMD CDNA™ 2 Architecture.



Source: AMD Instinct™ MI200 Series Accelerator and Node Architectures.

43. As shown above, when a CU core needs data/instructions that is not in its L1 cache, it makes a request. The request travels from the L1 cache to the L2 cache. If the data is not in the L2 cache, the request is sent to the HBM2e. This entire process is managed by the GCD's memory controllers. The In-package Infinity Fabric and/or internal Infinity Fabric handles the high-speed inter-node data communications between the two GCDs and their respective memory subsystems.

Data/instructions contained in the memory facilitates the inter-node data communications.

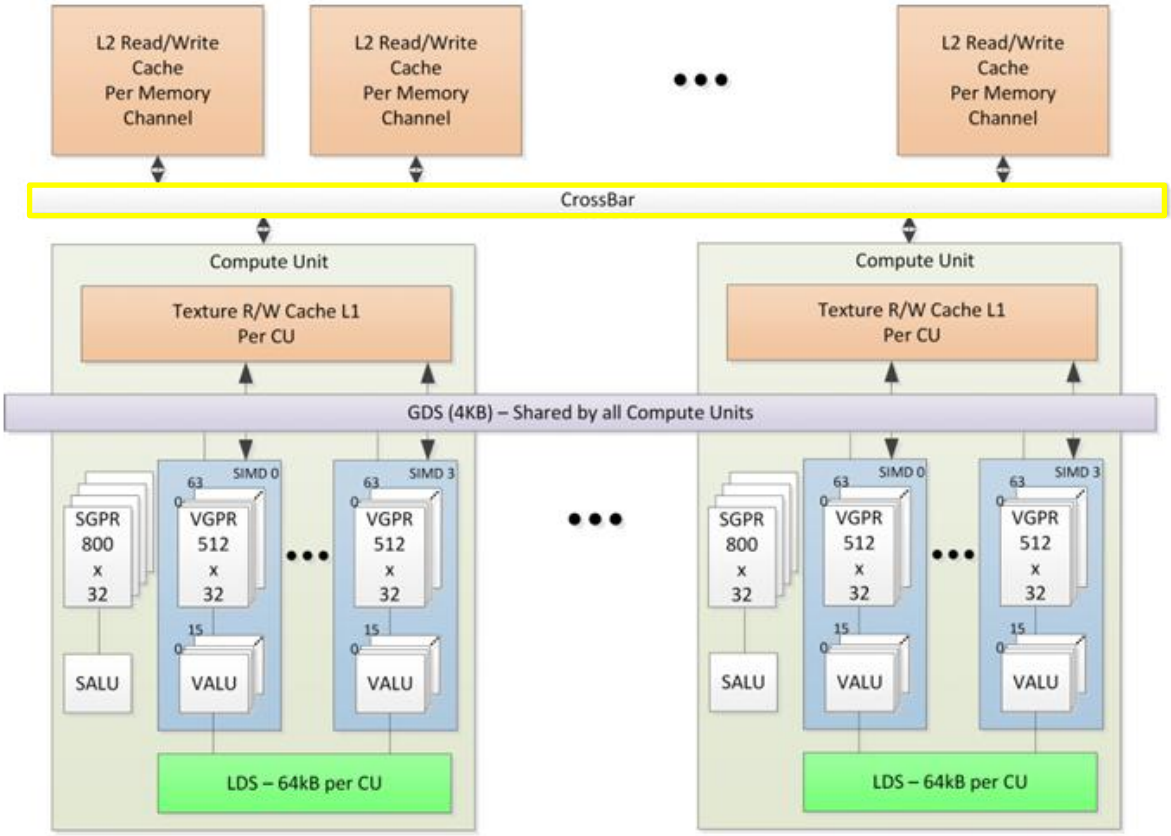


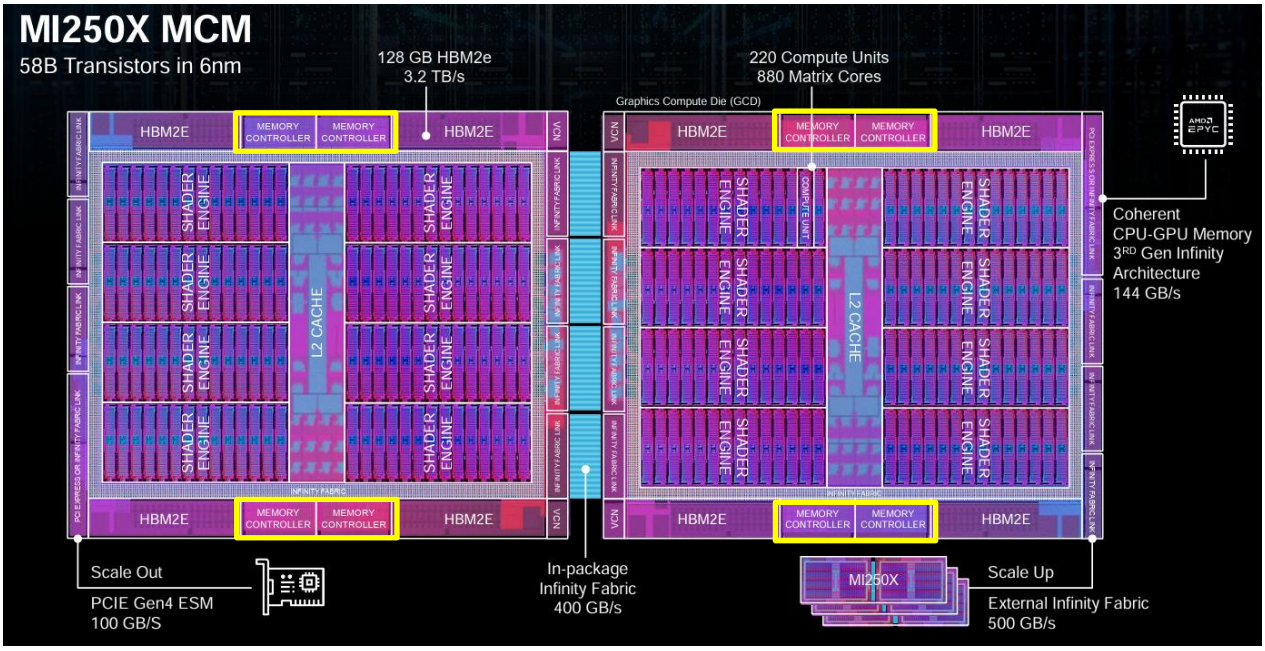
Figure 2. Shared Memory Hierarchy

Source: AMD Instinct MI200 Instruction Set Architecture.

44. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: a data mover component to facilitate intra-node data communications between the programmable logic component and the node memory at the respective compute node. As shown above, AMD Instinct™ MI200 Series Accelerator comprises a crossbar or bus ("a data mover component") (shown in the yellow box) to facilitate intra-node data communications between the CU/compute engine and the node memory (*e.g.,* an L2 R/W cache) at the respective GCD.

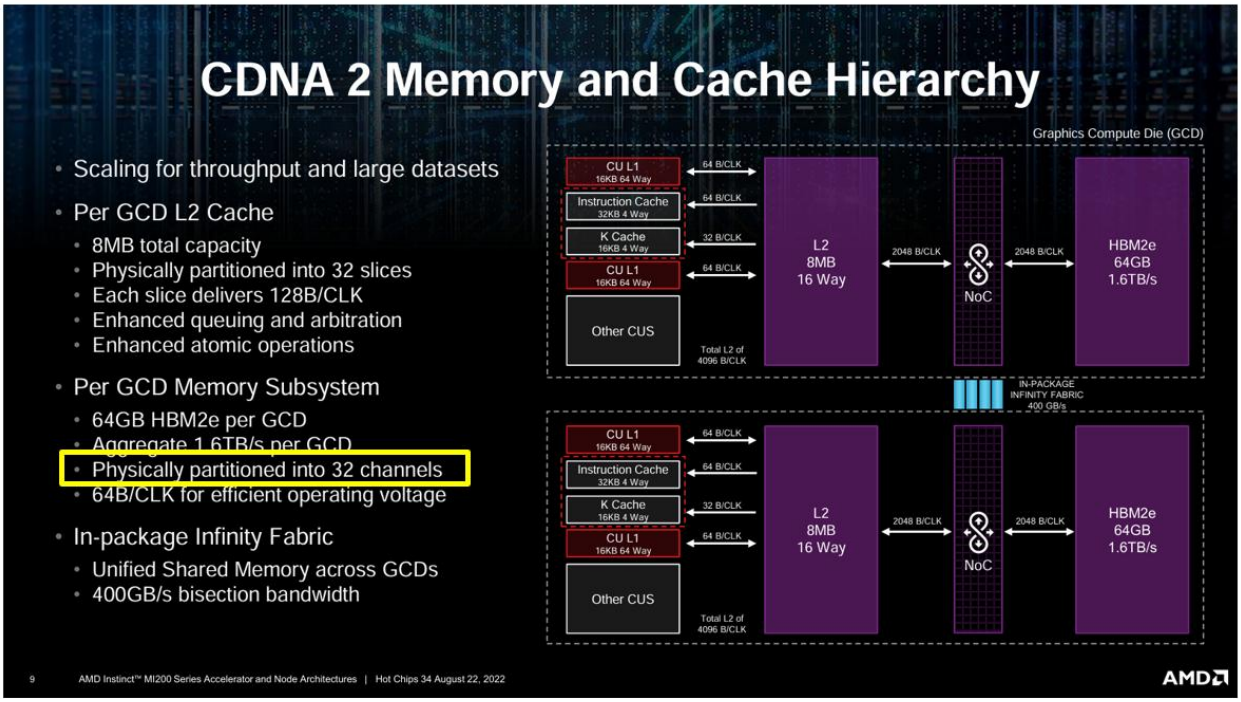
45. The AMD Instinct™ MI200 Series Accelerator is a data processing system

comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: a storage interface component to facilitate intra-node data communications between the programmable logic component and the data store at the respective compute node. Each GCD has a memory controller ("storage interface component") (shown in yellow boxes below) that transfers data intra-node from the HBM2e that is used by the GCD and CUs for data processing operations.



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

46. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a plurality of serially coupled compute nodes in an execution pipeline, each respective compute node in the execution pipeline comprising: wherein the storage interface component utilizes a plurality of communications channels to transfer data between the programmable logic component and the data store at the respective compute node. As shown above, the AMD Instinct™ MI200 Series Accelerator has 128GB of High Bandwidth Memory and 3.2TB/s bandwidth. As shown below, the GCD memory is partitioned into 32 channels.



Source: [AMD Instinct™ MI200 Series Accelerator and Node Architectures](#).

47. AMD is aware of the ’512 patent and the infringement of the ’512 patent by the Accused Products at least as of the filing of this Complaint.

48. On information and belief, at least as of the filing date of this Complaint, AMD has and continues to actively, knowingly, and intentionally induce infringement of at least claim 1 of the ’512 patent in violation of 35 U.S.C. § 271(b) and (f). AMD makes, uses, sells, offers for sale, and/or imports one or more of the Accused Products, or supplies or causes to be supplied in or from the United States all or a substantial portion of the components of the Accused Products, including the GCDs, for or on behalf of third parties such as customers, users, system builders, partners, and retailers/distributors, knowing and intending that the Accused Products will be used by the third parties in a manner that practices at least claim 1 of the ’512 patent as shown above.

49. For example, AMD publishes and provides marketing materials, technical

specifications, whitepapers, datasheets, user manuals, and development and testing information, and other resources on its website that instruct and encourage third parties to integrate the Accused Products into products using AMD's technology that were then made, used, sold, offered for sale and/or imported into the United States. AMD provides a website to inform customers the servers and workstations that support the Accused AMD Accelerator Products, case studies of customer success stories, and supporting documentation for the Instinct™ Accelerators.[1] These activities were designed to sell and bring infringing products that incorporate AMD's Accused Products to market in the United States. As a result of AMD's activities, the Accused Products have been used in a manner that directly infringes the '512 patent. AMD continues to engage in acts of inducement of infringement of the '512 patent.

50. AMD has also infringed, and continues to infringe, one or more claims of the '512 patent by selling, offering for sale, or importing into the United States, the Accused Products, knowing that the Accused Products constitute a material part of the inventions claimed in the '512 patent, are especially made or adapted to infringe the '512 patent, and are not staple articles or commodities of commerce suitable for non-infringing use. For example, the identified hardware and/or software components in AMD's Instinct™ MI200 Series Accelerators, including the GCDs, constitute a material part of the inventions claimed in the '512 patent, are especially made or

[1] *See, e.g.*, https://www.amd.com/content/dam/amd/en/documents/products/accelerators/instinct/amd-instinct-solution-catalog.pdf; https://www.amd.com/en/products/accelerators/instinct.html; https://www.amd.com/en/where-to-buy/accelerators/instinct.html#sortCriteria=%40amd_partner%20ascending; https://www.amd.com/en/resources/case-studies.html#sortCriteria=%40amd_release_date%20descending&f-amd_form_factor=Servers&f-amd_product_type=Accelerators; https://www.amd.com/en/search/documentation/hub.html#sortCriteria=%40amd_release_date%20descending&f-amd_product_type=Accelerators&f-amd_product_brand=Instinct&f-amd_product_series=Instinct%20MI200%20Series.

adapted to infringe the ’512 patent, and are not staple articles or commodities of commerce suitable for non-infringing use, as demonstrated by the evidence cited above. AMD has been, and currently is, contributorily infringing the ’512 patent in violation of 35 U.S.C. §§ 271(c) and (f).

51. Fermat has been damaged by the direct and indirect infringement of AMD and is suffering, and will continue to suffer, irreparable harm and damages as a result of that infringement. Fermat will suffer further irreparable injury, for which it has no adequate remedy at law, unless and until AMD is enjoined from infringing the claims of the ’512 patent.

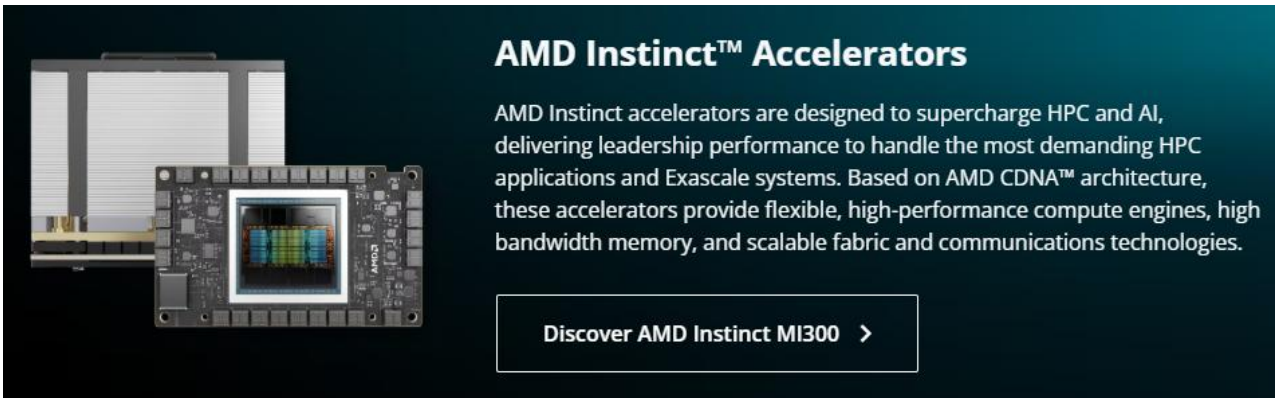
52. To the extent that AMD continues to infringe the ’512 patent after the filing of this Complaint, it does so in willful disregard of Fermat’s patent rights. AMD’s continued infringement of the ’512 patent is deliberate and willful, entitling Fermat to an award of treble damages, reasonable attorney fees, and costs in bringing this action.

Count II – Infringement of United States Patent No. 11,886,750

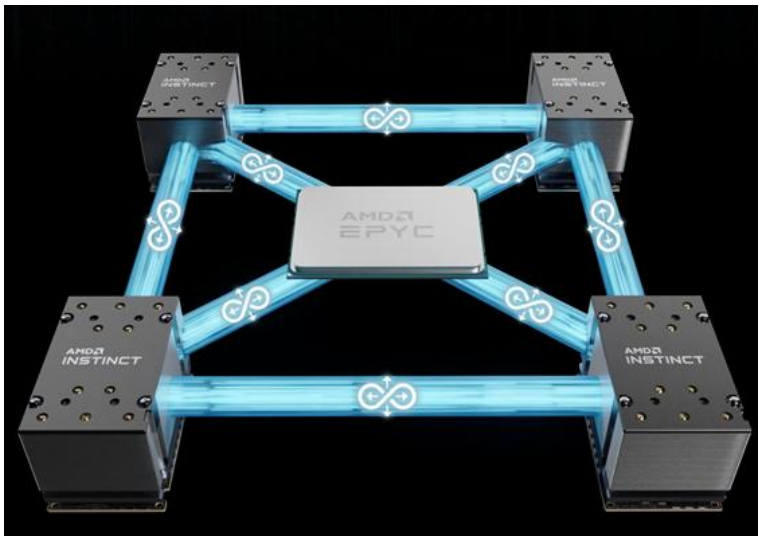
53. Fermat repeats, realleges, and incorporates by reference, as if fully set forth here, the allegations of the preceding paragraphs above.

54. AMD has directly infringed and continues to infringe, literally and/or through the doctrine of equivalents, at least claim 1 of the ’750 patent by making, using, selling, offering for sale, and/or importing one or more of the Accused Products, in violation of 35 U.S.C. § 271(a).

55. To the extent the preamble is limiting, the AMD Instinct™ MI200 Series Accelerator is a data processing system operative in cooperation with a host compute system. The AMD Instinct™ MI200 Series Accelerator is used for, *e.g.*, High Performance Computing (HPC) and Artificial Intelligence (AI) based applications, which involves various data processing operations. The AMD Instinct™ MI200 Series Accelerator operates in cooperation with the Central Processing Unit (“CPU”) (“host compute system”), such as AMD’s EPYC.



Source: HPC Solutions.



Source: AMD Instinct™ MI200 Series Accelerator and Node Architectures.

56. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising a router module comprising a host interface and a node interface, wherein the router module is communicatively coupled to the host compute system via the host interface. Fermat incorporates by reference paragraphs 32-34 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

57. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series. As discussed above, the AMD Instinct™ MI200 Series Accelerator is a multi-die

GPU. Pairs of GCDs ("compute nodes") in each AMD Instinct™ MI200 Series Accelerator are serially coupled using In-package Infinity Fabric ("first tier compute note coupled to one or more additional computer nodes in series"). Multiple parallel first tier compute nodes are connected in different topologies of MI200 Series Accelerators, shown in AMD's documentation (*e.g.*, Flagship ML Topology with MI250).

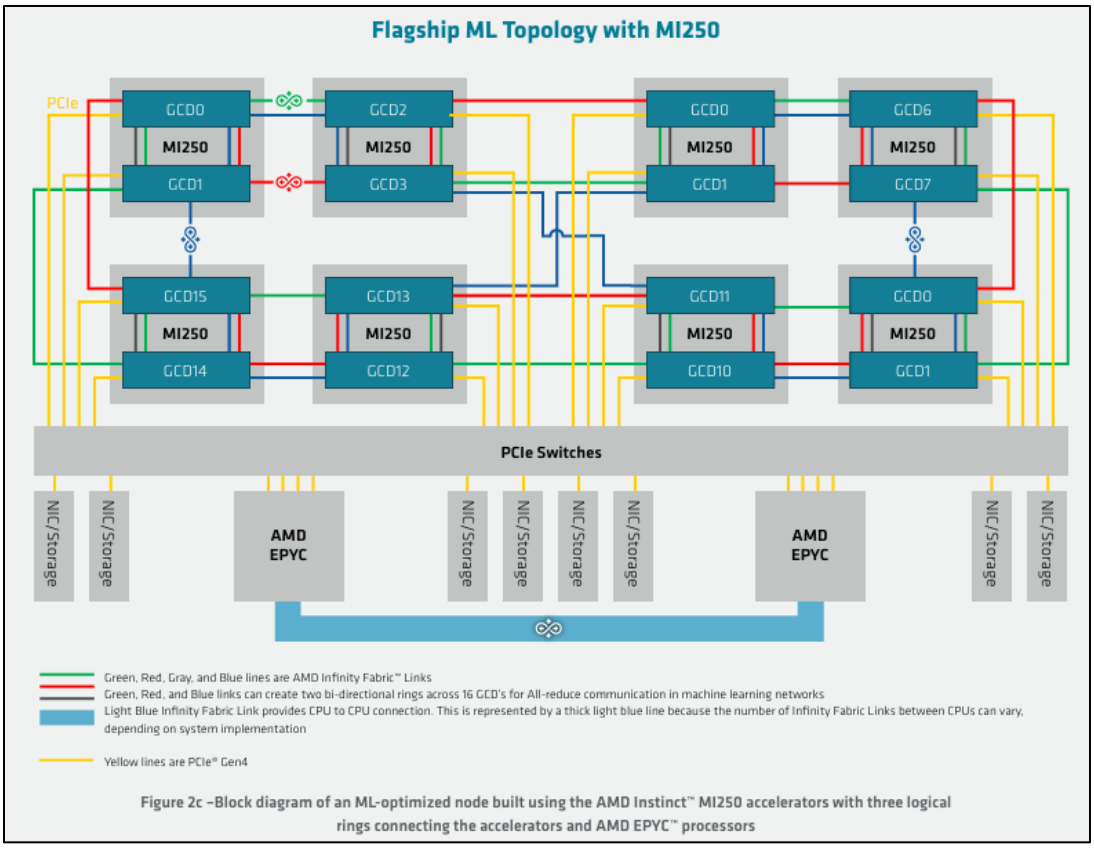


Figure 2c –Block diagram of an ML-optimized node built using the AMD Instinct™ MI250 accelerators with three logical rings connecting the accelerators and AMD EPYC™ processors

Source: AMD CDNA™ 2 ARCHITECTURE.

58. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a communications link, wherein each first tier compute node is communicatively coupled to the node interface at the router module via the communications link, at different successive tiers, each compute node in a series being communicatively coupled, via its communications link to one or more additional compute nodes

connected at different successive tiers.

59. For instance, the GCD0 and GCD1 are compute nodes connected in series as discussed in paragraph 35-. Further, GCD15 and GCD14 are at a different successive tier and are also in series and communicatively coupled by the Infinity Fabric communications link. Fermat incorporates by reference paragraph 35 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

60. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a data store comprising records associated with a data processing operation. Fermat incorporates by reference paragraph 38- above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

61. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a programmable logic component to execute the data processing operation in cooperation with the data store. Fermat incorporates by reference paragraph 39 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

62. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a node memory comprising data and instructions to support operation of the programmable logic component in connection with the data processing operation and to facilitate data communications via the communications link. Fermat incorporates by reference paragraphs 40-43 above discussing how the Accused Products infringe the ’512 patent

as if fully set forth herein.

63. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a data mover component to facilitate data communications between the programmable logic component and the node memory. Fermat incorporates by reference paragraph 44 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

64. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: a storage interface component to facilitate data communications between the programmable logic component and the data store. Fermat incorporates by reference paragraph 45 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

65. The AMD Instinct™ MI200 Series Accelerator is a data processing system comprising multiple parallel first tier compute nodes coupled to one or more additional compute nodes in series, each compute node comprising: wherein the storage interface component utilizes a plurality of communications channels to transfer data between the programmable logic component and the data store. Fermat incorporates by reference paragraph 46 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

66. AMD is aware of the ’750 patent and the infringement of the ’750 patent by the Accused Products at least as of the filing of this Complaint.

67. On information and belief, at least as of the filing date of this Complaint, AMD has and continues to actively, knowingly, and intentionally induce infringement of at least claim 1 of

the ’750 patent in violation of 35 U.S.C. § 271(b) and (f). AMD makes, uses, sells, offers for sale, and/or imports one or more of the Accused Products, or supplies or causes to be supplied in or from the United States all or a substantial portion of the components of the Accused Products, including the GCDs, for or on behalf of third parties such as customers, users, system builders, partners, and retailers/distributors, knowing and intending that the Accused Products will be used by the third parties in a manner that practices at least claim 1 of the ’750 patent as shown above.

68. For example, AMD publishes and provides marketing materials, technical specifications, whitepapers, datasheets, user manuals, and development and testing information, and other resources on its website that instruct and encourage third parties to integrate the Accused Products into products using AMD’s technology that were then made, used, sold, offered for sale and/or imported into the United States. AMD provides a website to inform customers the servers and workstations that support the Accused AMD Accelerator Products, case studies of customer success stories, and supporting documentation for the Instinct™ Accelerators.[2] These activities were designed to sell and bring infringing products that incorporate AMD’s Accused Products to market in the United States. As a result of AMD’s activities, the Accused Products have been used in a manner that directly infringes the ’750 patent. AMD continues to engage in acts of inducement of infringement of the ’750 patent.

---

[2] *See, e.g.*, https://www.amd.com/content/dam/amd/en/documents/products/accelerators/instinct/amd-instinct-solution-catalog.pdf; https://www.amd.com/en/products/accelerators/instinct.html; https://www.amd.com/en/where-to-buy/accelerators/instinct.html#sortCriteria=%40amd_partner%20ascending; https://www.amd.com/en/resources/case-studies.html#sortCriteria=%40amd_release_date%20descending&f-amd_form_factor=Servers&f-amd_product_type=Accelerators; https://www.amd.com/en/search/documentation/hub.html#sortCriteria=%40amd_release_date%20descending&f-amd_product_type=Accelerators&f-amd_product_brand=Instinct&f-amd_product_series=Instinct%20MI200%20Series.

69. AMD has also infringed, and continues to infringe, one or more claims of the ’750 patent by selling, offering for sale, or importing into the United States, the Accused Products, knowing that the Accused Products constitute a material part of the inventions claimed in the ’750 patent, are especially made or adapted to infringe the ’750 patent, and are not staple articles or commodities of commerce suitable for non-infringing use. For example, the identified hardware and/or software components in AMD’s Instinct™ MI200 Series Accelerators, including the GCDs, constitute a material part of the inventions claimed in the ’750 patent, are especially made or adapted to infringe the ’750 patent, and are not staple articles or commodities of commerce suitable for non-infringing use, as demonstrated by the evidence cited above. AMD has been, and currently is, contributorily infringing the ’750 patent in violation of 35 U.S.C. §§ 271(c) and (f).

70. Fermat has been damaged by the direct and indirect infringement of AMD and is suffering, and will continue to suffer, irreparable harm and damages as a result of that infringement. Fermat will suffer further irreparable injury, for which it has no adequate remedy at law, unless and until AMD is enjoined from infringing the claims of the ’750 patent.

71. To the extent that AMD continues to infringe the ’750 patent after the filing of this Complaint, it does so in willful disregard of Fermat’s patent rights. AMD’s continued infringement of the ’750 patent is deliberate and willful, entitling Fermat to an award of treble damages, reasonable attorney fees, and costs in bringing this action.

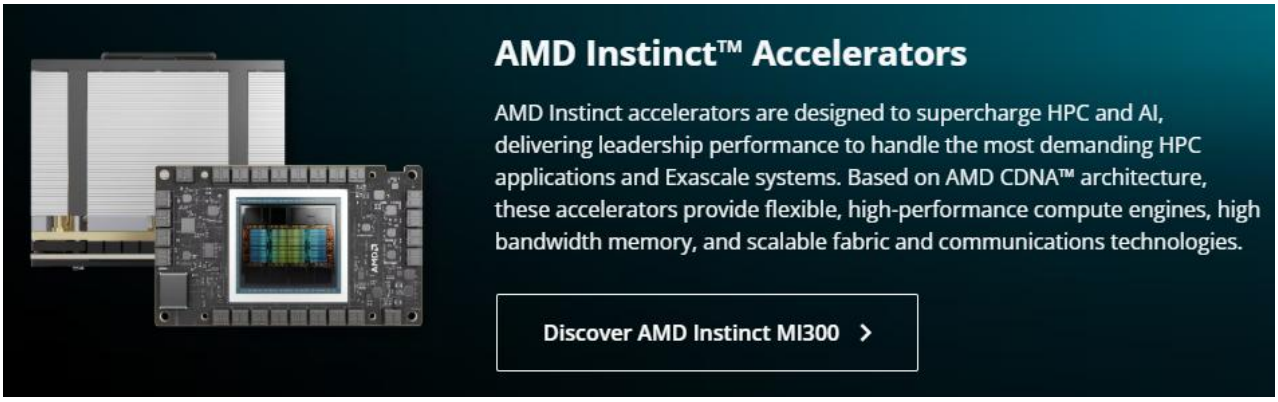
Count III – Infringement of United States Patent No. 12,271,636

72. Fermat repeats, realleges, and incorporates by reference, as if fully set forth here, the allegations of the preceding paragraphs above.

73. AMD has directly infringed and continues to infringe, literally and/or through the doctrine of equivalents, at least claim 1 of the ’636 patent by making, using, selling, offering for

sale, and/or importing one or more of the Accused Products, in violation of 35 U.S.C. § 271(a).

74. To the extent the preamble is limiting, the AMD Instinct™ MI200 Series Accelerator is a compute node for use in a data processing system. The AMD Instinct™ MI200 Series Accelerator is a multi-die GPU as shown above with a serially-coupled pair of GCDs (each a "compute node"). The AMD Instinct™ MI200 Series Accelerator is used for, *e.g.*, High Performance Computing (HPC) and Artificial Intelligence (AI) based applications, which involves various data processing operations.



Source: HPC Solutions.

75. The AMD Instinct™ MI200 Series Accelerator comprises a communications link communicatively coupling the compute node to one or both of a host compute system and an adjacent compute node in an execution pipeline. Fermat incorporates by reference paragraphs 36-37 above discussing how the Accused Products infringe the '512 patent as if fully set forth herein.

76. The AMD Instinct™ MI200 Series Accelerator comprises a data store comprising records associated with a data processing operation. Fermat incorporates by reference paragraph 38 above discussing how the Accused Products infringe the '512 patent as if fully set forth herein.

77. The AMD Instinct™ MI200 Series Accelerator comprises a programmable logic component to execute the data processing operation in cooperation with the data store. Fermat incorporates by reference paragraph 39 above discussing how the Accused Products infringe the

’512 patent as if fully set forth herein.

78. The AMD Instinct™ MI200 Series Accelerator comprises a node memory comprising data and instructions to support operation of the programmable logic component in connection with the data processing operation and to facilitate inter-node data communications with the adjacent node via the communications link. Fermat incorporates by reference paragraphs 40-43 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

79. The node memory further comprises algorithm memory including resource-specific memory locations (*e.g.,* for atomic operations) to support the data processing operation as shown in AMD’s documentation.

> Each GCD contains an L2 cache that is physically partitioned with one slice per memory controller and shared by all the resources on a single GCD. The AMD CDNA 2 family uses a 16-way set-associative design with 32 slices with a total capacity of 8MB (per GCD). To keep pace with the computational capabilities of the CUs, the bandwidth from each L2 slice has been doubled to 128B per clock – a peak of 6.96 TB/s for the MI250, more than 2x the prior generation[4]. The queuing and arbitration for the distributed L2 cache have been enhanced to improve utilization of this read bandwidth over a wide range of workloads.
>
> Not only did the L2 cache boost throughput, but the synchronization capabilities have significantly improved. Many algorithms, such as building histograms and computing other statistics, rely on atomic operations to coordinate communication across an entire GPU or even a cluster of GPUs. Some of these atomic operations are most naturally executed close to the memory in the L2 cache. The AMD CDNA 2 architecture boosts the throughput for FP64 atomic operations including addition, min, and max in the L2 cache.

Source: AMD CDNA2 White Paper.

> **9.5. Float Memory Atomics**
>
> Floating point memory atomics are executed in LDS and in the L2 cache. They can be issued as LDS, GDS, Buffer, Flat, Global, and Scratch instructions.
>
> This chapter explains the rules for rounding, denormals and NaN for floating point atomics.

Source: "AMD Instinct MI200" Instruction Set Architecture.

80. The AMD Instinct™ MI200 Series Accelerator comprises a data mover component to facilitate intra-node data communications between the programmable logic component and the node memory. Fermat incorporates by reference paragraph 44 above discussing how the Accused

Products infringe the ’512 patent as if fully set forth herein.

81. The AMD Instinct™ MI200 Series Accelerator comprises a storage interface component to facilitate intra-node data communications between the programmable logic component and the data store. Fermat incorporates by reference paragraph 45 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

82. The AMD Instinct™ MI200 Series Accelerator comprises wherein the storage interface component utilizes a plurality of communications channels to transfer data between the programmable logic component and the data store. Fermat incorporates by reference paragraph 46 above discussing how the Accused Products infringe the ’512 patent as if fully set forth herein.

83. The AMD Instinct™ MI200 Series Accelerator comprises wherein the data processing operation is associated with a computationally expensive algorithm. The AMD Instinct™ MI200 Series Accelerator GCD’s execute high performance computing (HPC) algorithms, AI algorithms, and other computationally intensive applications operation (*e.g.*, data processing operations associated with a computationally expensive algorithm).

AMD INSTINCT™ MI200 SERIES ACCELERATOR

World’s Fastest HPC and AI Accelerator[1]

The era of exascale is here. Immense computational power coupled with the fusion of HPC and AI is enabling researchers and scientists to tackle our most pressing challenges from climate change to vaccine research. With the AMD Instinct™ MI200 accelerators and ROCm™ 5.0 software ecosystem, innovators can tap the power of the world’s most powerful HPC and AI data center GPUs to accelerate their time to science and discovery.[1]

Based on the 2nd Gen AMD CDNA™ architecture, AMD Instinct™ MI200 accelerators deliver a quantum leap in HPC and AI performance over competitive data center GPUs today. With an up to 4x advantage in HPC performance compared to competitive GPUs, the MI200 accelerator is the first data center GPU to deliver 383 teraflops of theoretical mixed precision FP16 performance for deep learning training, offering users a powerful platform to fuel the convergence of HPC and AI.[1]

Source: AMD INSTINCT™ MI200 SERIES ACCELERATOR.

84. The GCD "compute node" of the AMD Instinct™ MI200 Series Accelerator is deployed in a distributed processing compute environment to execute the computationally expensive algorithm, including with the AMD EPYC CPU. As shown above, the GCDs are connected with each other through in-package Infinity Fabric and the EPYC, and the GCDs are designed to provide a parallel micro-architecture for computationally intensive data applications.

> **Chapter 1. Introduction**
>
> AMD CDNA processors implement a parallel micro-architecture that is designed to provide an excellent platform for general-purpose data parallel applications. Data-intensive applications that require high bandwidth or are computationally intensive are a candidate for running on an AMD CDNA processor.

Source: AMD Instinct MI200 Instruction Set Architecture.

85. AMD is aware of the '636 patent and the infringement of the '636 patent by the Accused Products at least as of the filing of this Complaint.

86. On information and belief, at least as of the filing date of this Complaint, AMD has and continues to actively, knowingly, and intentionally induce infringement of at least claim 1 of the '636 patent in violation of 35 U.S.C. § 271(b) and (f). AMD makes, uses, sells, offers for sale, and/or imports one or more of the Accused Products, or supplies or causes to be supplied in or from the United States all or a substantial portion of the components of the Accused Products, including the GCDs, for or on behalf of third parties such as customers, users, system builders, partners, and retailers/distributors, knowing and intending that the Accused Products will be used by the third parties in a manner that practices at least claim 1 of the '636 patent as shown above.

87. For example, AMD publishes and provides marketing materials, technical specifications, whitepapers, datasheets, user manuals, and development and testing information, and other resources on its website that instruct and encourage third parties to integrate the Accused

Products into products using AMD's technology that were then made, used, sold, offered for sale and/or imported into the United States. AMD provides a website to inform customers the servers and workstations that support the Accused AMD Accelerator Products, case studies of customer success stories, and supporting documentation for the Instinct™ Accelerators.[3] These activities were designed to sell and bring infringing products that incorporate AMD's Accused Products to market in the United States. As a result of AMD's activities, the Accused Products have been used in a manner that directly infringes the '636 patent. AMD continues to engage in acts of inducement of infringement of the '636 patent.

88. AMD has also infringed, and continues to infringe, one or more claims of the '636 patent by selling, offering for sale, or importing into the United States, the Accused Products, knowing that the Accused Products constitute a material part of the inventions claimed in the '636 patent, are especially made or adapted to infringe the '636 patent, and are not staple articles or commodities of commerce suitable for non-infringing use. For example, the identified hardware and/or software components in AMD's Instinct™ MI200 Series Accelerators, including the GCDs, constitute a material part of the inventions claimed in the '636 patent, are especially made or adapted to infringe the '636 patent, and are not staple articles or commodities of commerce suitable for non-infringing use, as demonstrated by the evidence cited above. AMD has been, and currently

[3] *See, e.g.*, https://www.amd.com/content/dam/amd/en/documents/products/accelerators/instinct/amd-instinct-solution-catalog.pdf; https://www.amd.com/en/products/accelerators/instinct.html; https://www.amd.com/en/where-to-buy/accelerators/instinct.html#sortCriteria=%40amd_partner%20ascending; https://www.amd.com/en/resources/case-studies.html#sortCriteria=%40amd_release_date%20descending&f-amd_form_factor=Servers&f-amd_product_type=Accelerators; https://www.amd.com/en/search/documentation/hub.html#sortCriteria=%40amd_release_date%20descending&f-amd_product_type=Accelerators&f-amd_product_brand=Instinct&f-amd_product_series=Instinct%20MI200%20Series.

is, contributorily infringing the ’636 patent in violation of 35 U.S.C. §§ 271(c) and (f).

89. Fermat has been damaged by the direct and indirect infringement of AMD and is suffering, and will continue to suffer, irreparable harm and damages as a result of that infringement. Fermat will suffer further irreparable injury, for which it has no adequate remedy at law, unless and until AMD is enjoined from infringing the claims of the ’636 patent.

90. To the extent that AMD continues to infringe the ’636 patent after the filing of this Complaint, it does so in willful disregard of Fermat’s patent rights. AMD’s continued infringement of the ’636 patent is deliberate and willful, entitling Fermat to an award of treble damages, reasonable attorney fees, and costs in bringing this action.

PRAYER FOR RELIEF

Fermat prays for judgment in its favor against AMD for the following relief:

A. Pursuant to 35 U.S.C. § 271, a determination that AMD and its officers, agents, employees, and all others in active concert and/or participation with them have infringed, directly and indirectly, each of the ’512, ’750, and ’636 patents through the manufacture, use, testing, importation, offer for sale, and/or sale of infringing products and/or components and/or any of the other acts prohibited by 35 U.S.C. § 271;

B. Pursuant to 35 U.S.C. § 283, an injunction enjoining AMD and its officers, agents, employees, and all others in active concert and/or participation with them from further infringing the ’ 512, ’750, and ’636 patents through the manufacture, use, testing, sale, offer for sale, importation and/or any of the other acts prohibited by 35 U.S.C. § 271, including preliminary and permanent injunctive relief;

C. Pursuant to 35 U.S.C. § 284, an award compensating Fermat for AMD’s infringement of the ’512, ’750, and ’636 patents through payment of not less than a reasonable

royalty on AMD's sales of infringing products, including an accounting for supplemental damages through judgment and a compulsory license for ongoing infringement;

D. A judgment and order finding that AMD's acts of infringement were egregious and willful and trebling damages under 35 U.S.C. § 284;

E. Pursuant to 35 U.S.C. § 285, a finding that this is an exceptional case, and an award of reasonable attorneys' fees and non-taxable costs;

F. An assessment of prejudgment and post-judgment interest and costs against AMD, together with an award of such interest and costs, pursuant to 35 U.S.C. § 284;

G. That this Court award such other and further relief as this Court may deem just.

DEMAND FOR A JURY TRIAL

Pursuant to Federal Rule of Civil Procedure 38(b), Fermat hereby demands a trial by jury of all issues so triable.

Dated: January 13, 2026

Respectfully submitted,

/s/ Elizabeth Bernard
Scott R. Samay* (*pro hac vice to be filed*)
Jason S. Charkow* (*pro hac vice to be filed*)
Elizabeth Bernard (admitted DC Bar No. 986675)
Stephanie Mandir* (*pro hac vice to be filed*)
Taylor Lepore* (*pro hac vice to be filed*)
Lisa Phillips* (*pro hac vice to be filed*)
Raymond W. Mort, III (TX State Bar No. 00791308)
ssamay@daignaultiyer.com
jcharkow@daignaultiyer.com
ebernard@daignaultiyer.com
smandir@daignaultiyer.com
tlepore@daignaultiyer.com
lphillips@daignaultiyer.com
rmort@daignaultiyer.com
DAIGNAULT IYER LLP
8229 Boone Boulevard, Suite 450
Vienna, VA 22182

Attorneys for Plaintiff Fermat International, Inc.