INTRODUCING

AMD CDNA™ 2 ARCHITECTURE

Propelling humanity's foremost research with the world's most powerful HPC and AI accelerator.

AMD INSTINCT | AMD ROCm

AMD

Table of Contents

Introduction

The history of GPUs is a story of evolution - from extremely specialized and hardwired designs to fully programmable accelerators. These accelerators are fundamentally optimized for massive throughput, and can run standard programming languages like C++ as part of a rich software ecosystem. This combination has proven incredibly compelling for scientific computing and machine learning, where the computational throughput has enabled tremendous innovation and discovery in a wide variety of applications.

Prior accelerator architectures have consistently improved performance and efficiency while become increasingly programmable. However, the AMD CDNA™ 2 architecture takes this evolutionary path to the next level, achieving over 4x[1] performance boost over the prior generation AMD CDNA architecture, with 47.9 TFLOP/s peak double-precision vector FP64 throughput, to enable exascale levels of performance with unrivalled programmability in heterogeneous systems.

The AMD CDNA 2 architecture represents a major leap forward compared to the prior generation by enhancing the Matrix Core technology for HPC and AI, driving computational capabilities for double-precision floating-point data and a variety of matrix-multiply primitives. It also focuses on improving accelerator communication and scaling, leveraging AMD's unique Infinity Fabric™ to enable a multi-chip module that provides 1.8x the compute density[2] over the previous generation and enables better connectivity within a single system. Lastly, the AMD CDNA 2 architecture enables accelerators such as the AMD Instinct™ MI250X to operate as a full peer within a computing system by offering cache coherency with select optimized 3rd Gen EPYC processors, offering a quick and simple on-ramp for CPU codes to tap the power of accelerators. These new capabilities are all seamlessly unlocked by AMD's ROCm™ open software platform, which offers a rich set of tools for customers porting or developing cutting edge applications.

AMD CDNA™ 2 Architecture Overview

The AMD CDNA™ 2 architecture builds on the tremendous core strengths of the original AMD CDNA architecture to deliver a leap forward in system performance and usability while using a similar process technology. The AMD CDNA architecture is an excellent starting point for a computational platform. However, to deliver exascale performance the architecture was overhauled with enhancements to nearly every aspect from the compute units to the memory interface, with a particular emphasis on radically improving the communication interfaces for full system scalability.

The AMD CDNA 2 architecture is designed first and foremost for the most taxing scientific computing and machine learning applications. It powers the new AMD Instinct™ MI200 generation of products that target solutions ranging from compact single systems all the way to the world's largest exascale supercomputers with unique and highly differentiated programming models. Figure 1a illustrates the AMD Instinct MI200 Graphics Compute Die (GCD). The AMD Instinct™ MI200 is built on advanced packaging technologies, enabling two GCDs to be integrated into a single package in the OAM (OCP Accelerator Module) form factor in the MI250 and MI250X products, as shown in figure 1b. The AMD Instinct MI210 product enables one GCD in a traditional PCIe® form factor card for mainstream HPC and AI workloads in the data center. Each GCD is built on the AMD CDNA 2 architecture.

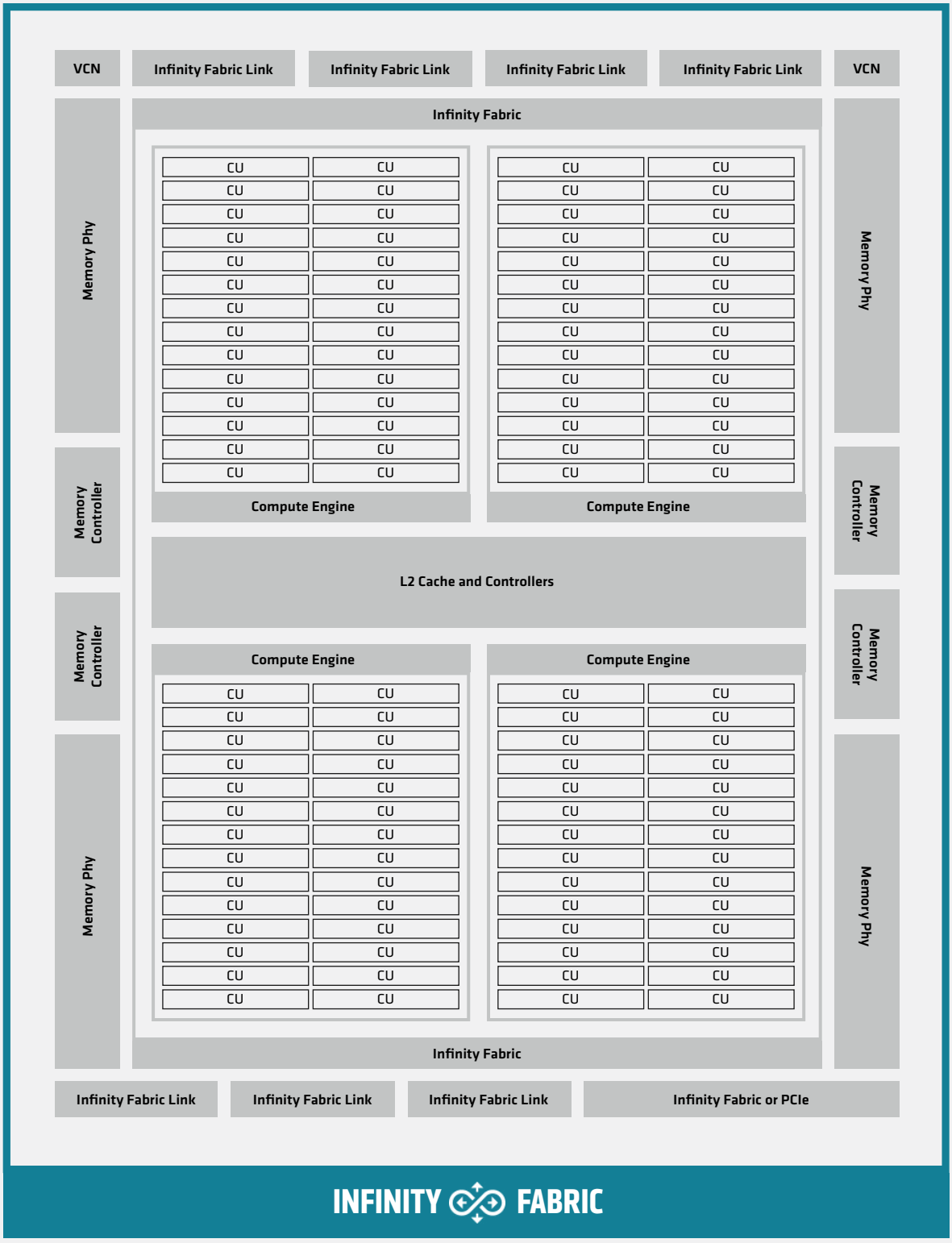


Figure 1a –Block diagram of the AMD Instinct™ MI200 Graphics Compute Die (GCD)

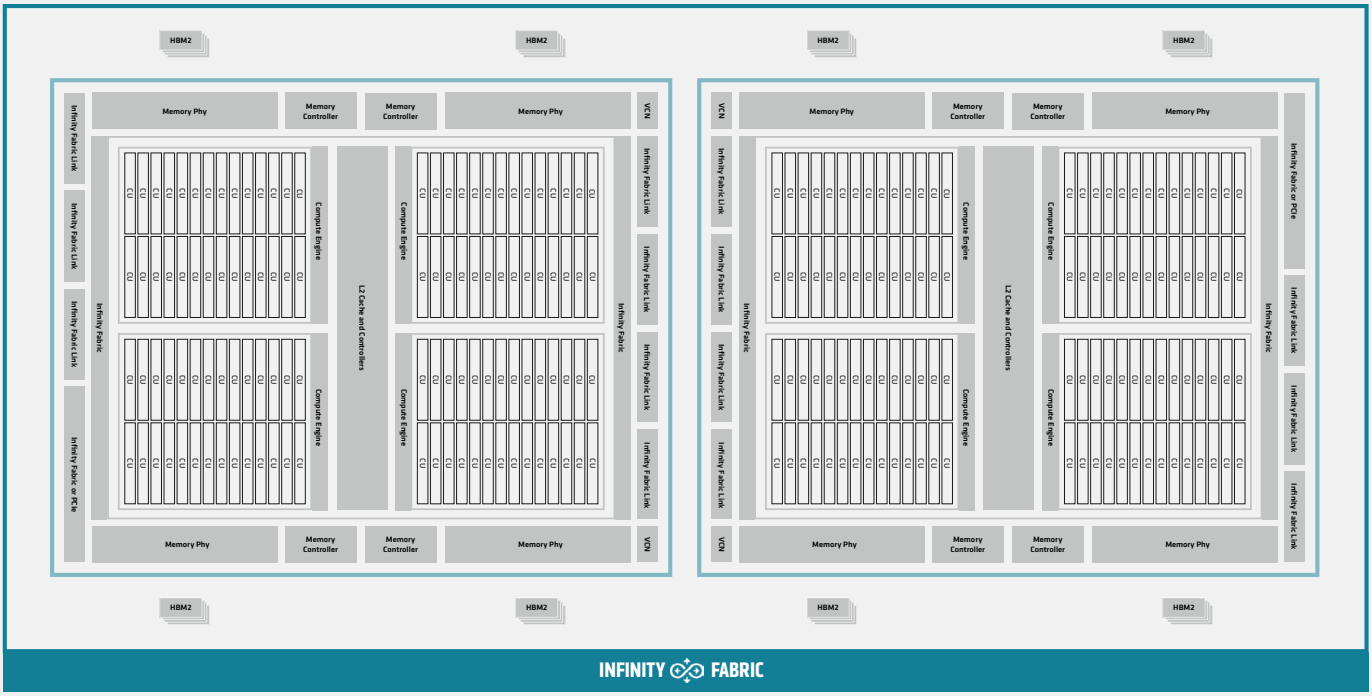


Figure 1b –Block diagram of the AMD Instinct™ MI200 multi-chip module (AMD Instinct™ MI250/MI250X) in the OAM form factor accelerator which comprise two Graphics Compute Dies (GCD) as illustrated.

The three critical functions of any processor are compute, memory, and communication. Each AMD CDNA™ 2 GCD has multiple blocks dedicated to these functions and are connected with a high-speed on-die fabric. However, to deliver exascale-performance this alone is not enough. One of the crucial innovations in the AMD CDNA 2 architecture is employing AMD's unique Infinity Fabric to extend the on-die fabric across the package so that each GCD appears as a GPU in one shared memory system. Connecting two GCDs together in this fashion doubles the resources, creating a larger computational building block on top of the many other enhancements.

The AMD CDNA 2 architecture has several different incarnations offering both a custom implementation that uses AMD Infinity Fabric™ to interface with an optimized 3rd generation AMD EPYC™ processor for a specific HPE/Cray supercomputer platform as well as a generally available implementation that relies on PCI-Express® to interface with host processors. Each GCD includes a command processor that gets API-level commands from the host CPU and translates them into work that can be spawned on different parts of the AMD CDNA 2 architecture.

One of the fundamental innovations in the prior generation AMD CDNA architecture was the introduction of the Matrix Core technology in the compute units (CUs) to boost computational throughput with a focus on datatypes used in machine learning. The Matrix Core technology in the AMD CDNA 2 architecture builds on this foundation and has been enhanced to support a wider range of datatypes and applications, with a particular emphasis on scientific computing with FP64 data. Additionally, similar to the prior generation, the CU array is partitioned into four shader engines that execute the compute kernels that are spawned by the command processor. The net result is that AMD Instinct MI200 series accelerators can provide up to a theoretical peak 47.9 TFLOP/s double-precision throughput, 4.2X the previous generation[1].

Each GCD also features 2x Video Codec Next (VCN), a logic block that provides encode and decode capabilities on incoming and outgoing data streams. This is a particularly useful logic for Machine Learning Training workloads for object detection that train on image and video data. The VCN blocks support H.264/AVC, HEVC, VP9 and JPEG for decode, as well as H.264/AVC and HEVC for encode[3].

Scaling the Memory Hierarchy

The AMD CDNA™ 2 architecture boosts the computational throughput compared to the prior generation architecture. The extremely high bandwidth register files and local data storage are tailored to support this performance. Additionally, the shared memory hierarchy outside of the CUs is critical to delivering the bandwidth necessary for real-world applications which work on large scale datasets that reside in memory. The AMD CDNA 2 memory controllers focus on efficiently accessing large working sets of data and bringing it on-die so that the L2 cache can provide amplified the bandwidth to feed the CUs. The AMD CDNA 2 architecture boosts many different dimensions of the memory hierarchy, simultaneously improving bandwidth[5] generationally and capacity while enhancing synchronization. Additionally, for select optimized system implementations in platforms utilizing the flagship HPC topology shown in figure 2a, it offers a unique coherent memory model that is built on top of AMD Infinity Fabric™.

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.

The AMD CDNA 2 memory capabilities have been scaled up in tandem with the computational requirements to handle large exascale-class computing problems. The memory capacity has doubled to 64GB per GCD from 32GB in the prior generation, so that a multichip MI200 series accelerator (MI250, MI250X) can access up to 128GB of data (which is 64GB per GCD x 2 per AMD Instinct MI250/250X device) – 4x the total memory capacity than the prior generation[5] and comparable to the main memory of an entire server from a decade ago. The HBM2e memory interface operates at aggregate 3.2TB/s which is 2.7x the previous generation[5] of peak theoretical memory bandwidth taking into account the dual GCDs. The AMD Instinct MI210 accelerators provide up to 64GB High-bandwidth HBM2e memory at a clock rate of 1.6GHz and offer an ultra-high 1.6 TB/s of memory bandwidth[10]. To keep pace with this increase in off-chip bandwidth, the connection between the individual memory controllers and L2 cache slices has doubled to 64-bytes wide.

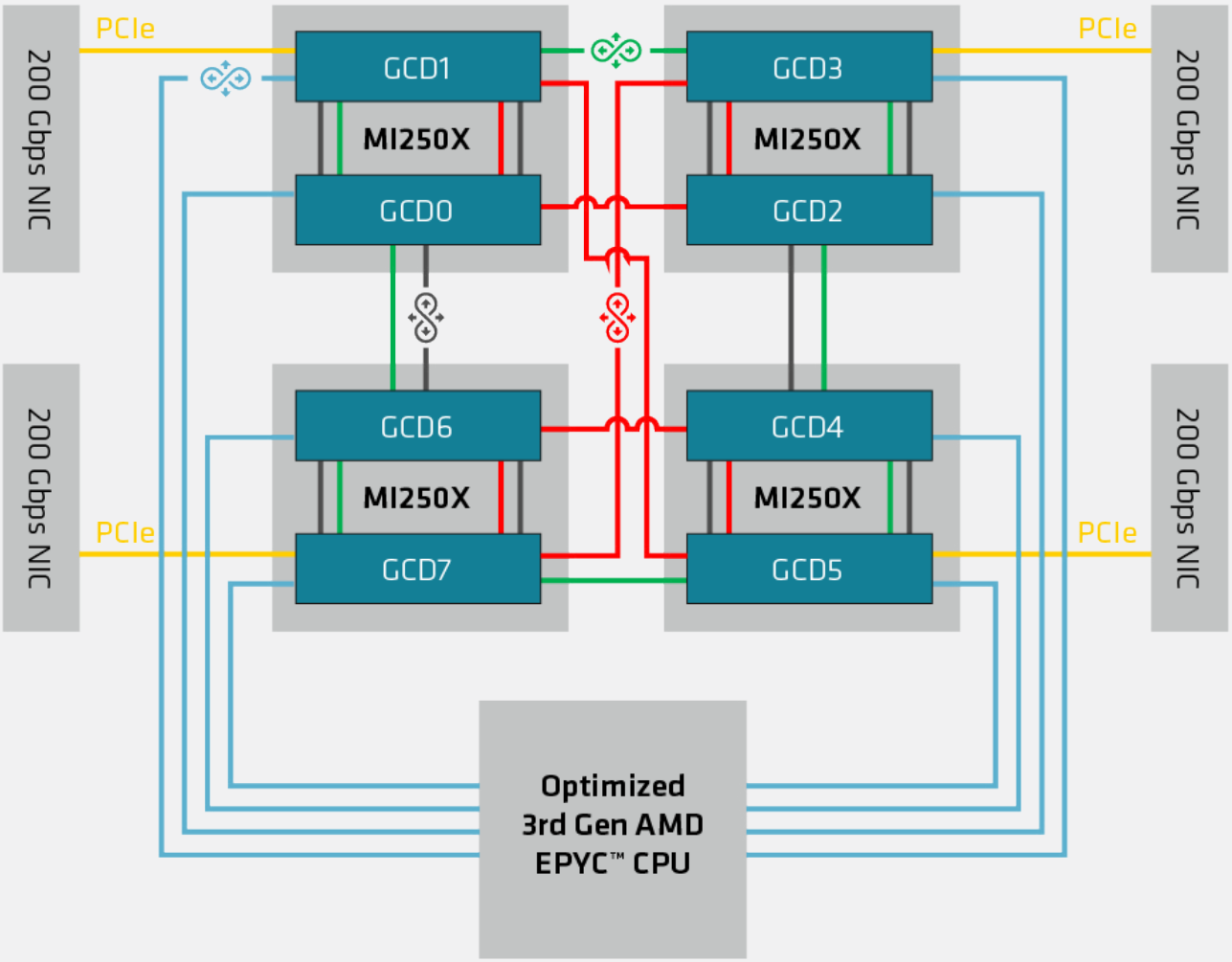
Beyond these quantitative improvements, the AMD CDNA 2 architecture also features an advanced and improved coherency model when used with the optimized 3rd gen AMD EPYC™ processor. This unique and qualitatively better coordination built around the Infinity Fabric supports coherent CPU-GPU memory to tackle the largest and most demanding problems. In this case, the AMD CDNA 2 architecture allows the optimized 3rd gen AMD EPYC CPU to cache the GPU memory with invalidation from both sides. This enables simplifying many workloads such as GROMACS[5] for molecular dynamics or the HACC[6] cosmological simulator. The GPU maintains a directory that tracks what portions of memory are shared with the CPU to avoid performance overhead. The GPU memory addressing has also expanded to satisfy large scale systems with up to 4PB of physically addressable memory and 57-bits or 128 petabytes of virtual memory, which is critical for memory coherency with the optimized 3rd gen AMD EPYC CPU.

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.

The MI210 accelerator with a single GCD also benefits from AMD Infinity Fabric technology. AMD Instinct MI210 GPUs can provide advanced I/O capabilities in standard off-the-shelf servers with external Infinity Fabric and PCIe® Gen4 support with bridge boards attached to hives of GPUs for high-speed GPU to GPU communication. The MI210 GPU delivers 64 GB/s CPU to GPU bandwidth without the need for PCIe® switches, and offers up to 300 GB/s of Peer-to-Peer (P2P) bandwidth performance through three Infinity Fabric links. The AMD Infinity Architecture enables platform designs with two GPU and four GPU direct-connect hives, using high-speed P2P connectivity bridge boards and offer up to 1.2 TB/s of total theoretical GPU bandwidth within a server design11 (Figure 2d and 2e)

Flagship HPC Topology with MI250X

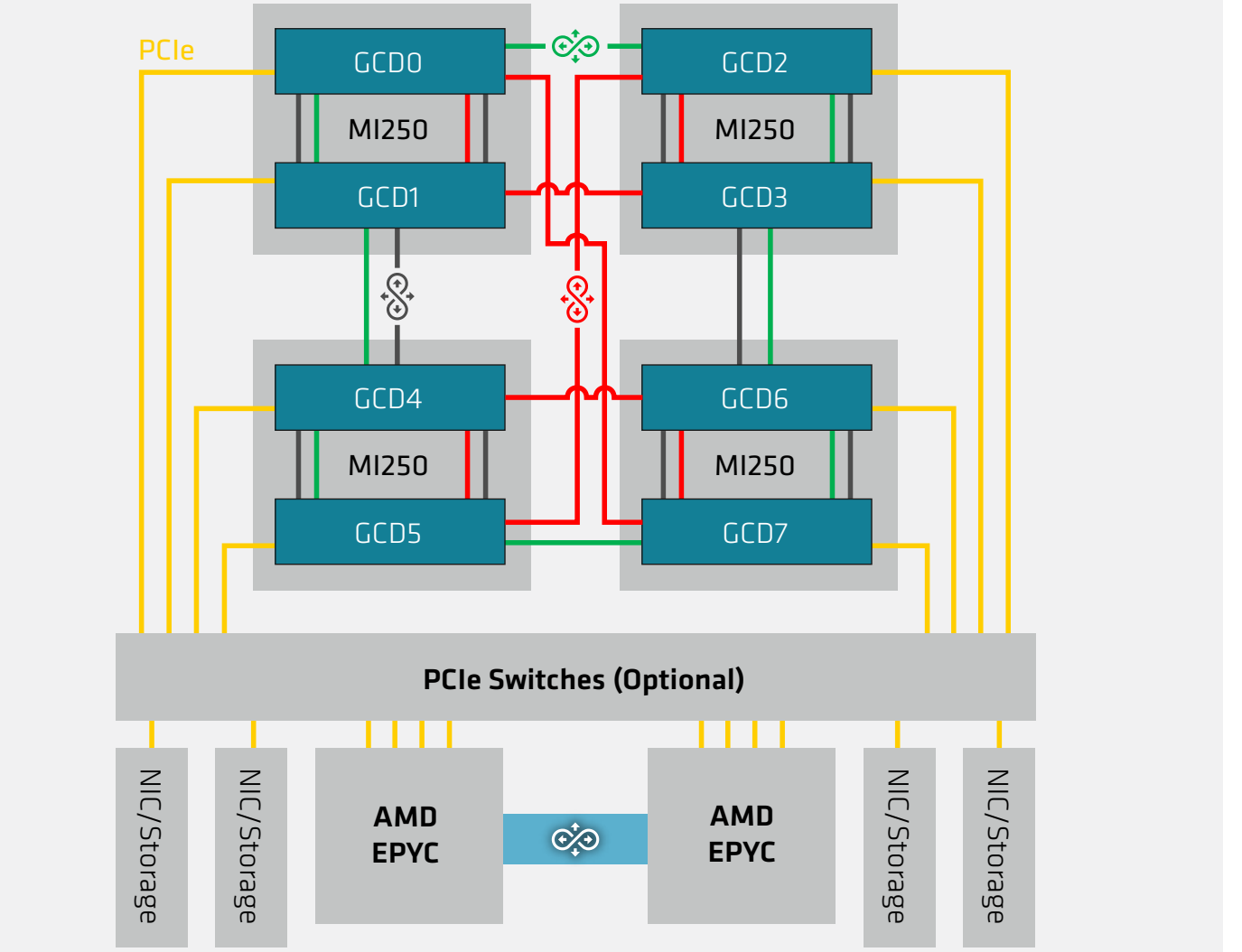


Green, Red, Gray, and Blue lines are AMD Infinity Fabric™ Links
Red and Green links can create two bi-directional rings
Blue Infinity Fabric Link provides coherent GCD-CPU connection

Orange lines are PCIe® Gen4 with ESM

Figure 2a – Block diagram of a flagship HPC node built using the AMD Instinct™ MI250X accelerator and optimized 3rd generation AMD EPYC™ processor

Mainstream HPC/ML Topology with MI250X



Green, Red, Gray, and Blue lines are AMD Infinity Fabric™ Links
Red and Green links can create two bi-directional rings for All-reduce communication in machine learning networks
Blue Infinity Fabric Link provides CPU to CPU connection. This is represented by a thick blue line because the number o
Infinity Fabric Links between CPUs can vary, depending on system implementation

Yellow lines are PCIe® Gen4

Figure 2b – Block diagram of a mainstream HPC/ML node built using the AMD Instinct™ MI250 accelerators and AMD EPYC™ processors

Flagship ML Topology with MI250

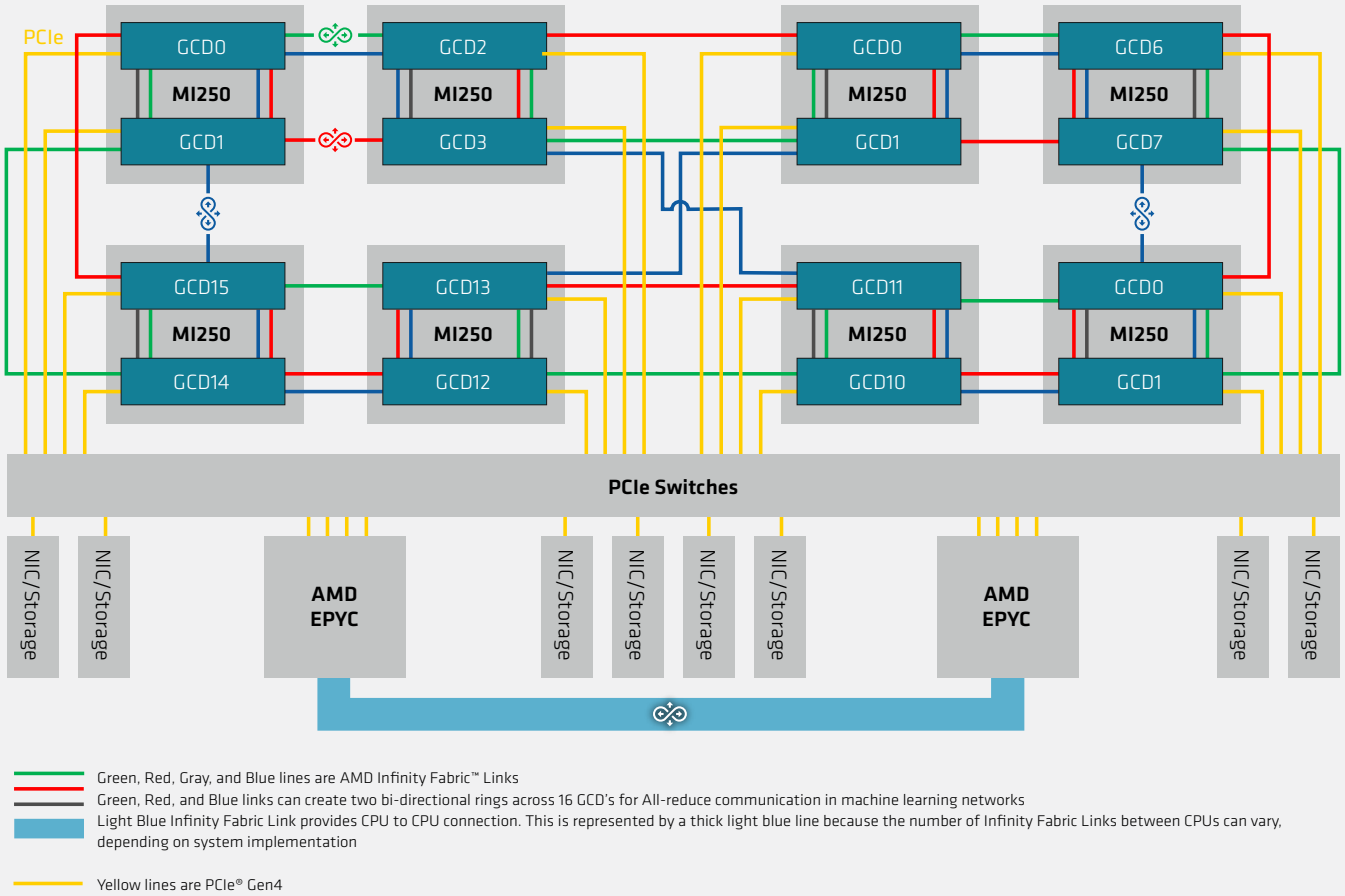


Figure 2c –Block diagram of an ML-optimized node built using the AMD Instinct™ MI250 accelerators with three logical rings connecting the accelerators and AMD EPYC™ processors

The in-package AMD Infinity Fabric™ interface is one of the key innovations in the AMD CDNA™ 2 family, connecting the 2 GCDs within the MI250 or MI250X. It takes advantage of the extremely short distances between the GCDs within the package to operate at 25 Gbps and at extremely low power, delivering a theoretical maximum bi-directional bandwidth of up to 400 GB/s[7] between the GCDs.

The 8 external AMD Infinity Fabric™ links for GPU P2P or I/O on the AMD Instinct™ MI250 (or MI250X) accelerators delivers up to 800 GB/s of total theoretical bandwidth providing up to 235% the GPU P2P (or I/O) theoretical bandwidth performance of the previous generation AMD Instinct™ GPU compute products[9].

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.

The last interface is a downstream PCIe 4.0 ESM link that operates at up to 25Gbps. Unlike the host interface, this downstream interface is coupled to a PCIe root complex, which can drive I/O devices that are connected to the GPU. This capability is crucial for the vision of exascale computing, which relies on CPUs and GPUs acting as equals in a system. With the introduction of full coherency between the CPU and the GPU, these two devices will act as peers for computation. This downstream I/O link allows them to both connect to high-speed networking and act as full peers in the context of communication.

Mainstream MI210 HPC/ML Topology

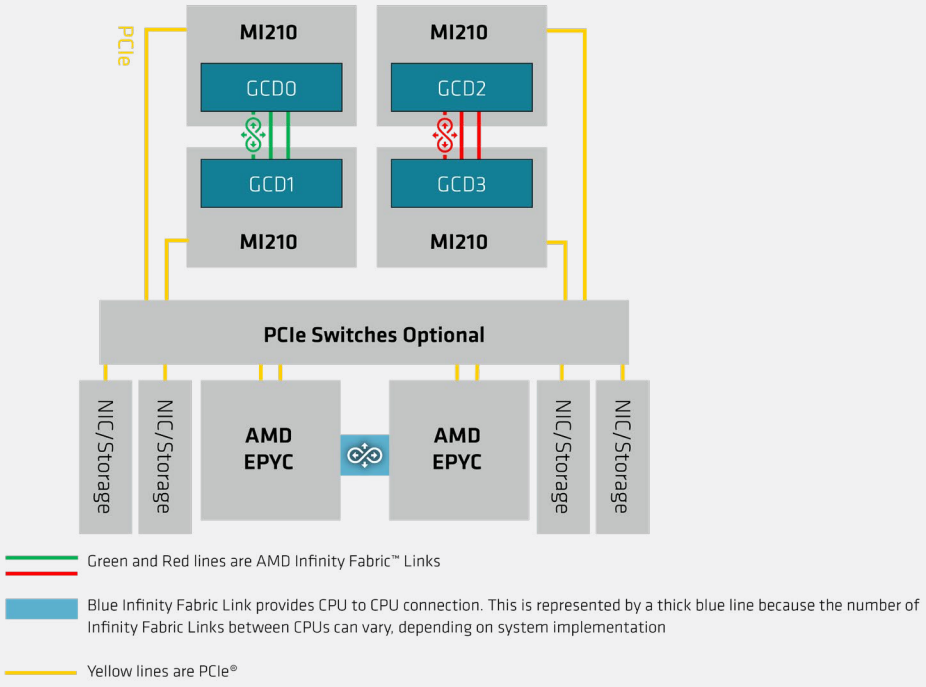


Figure 2d - Block diagram of a mainstream HPC/AI server node built using the AMD Instinct™ MI210 accelerators in a dual GPU hive with direct connect xGMI bridge boards and AMD EPYC™ processor

Flagship MI210 HPC/ML Topology

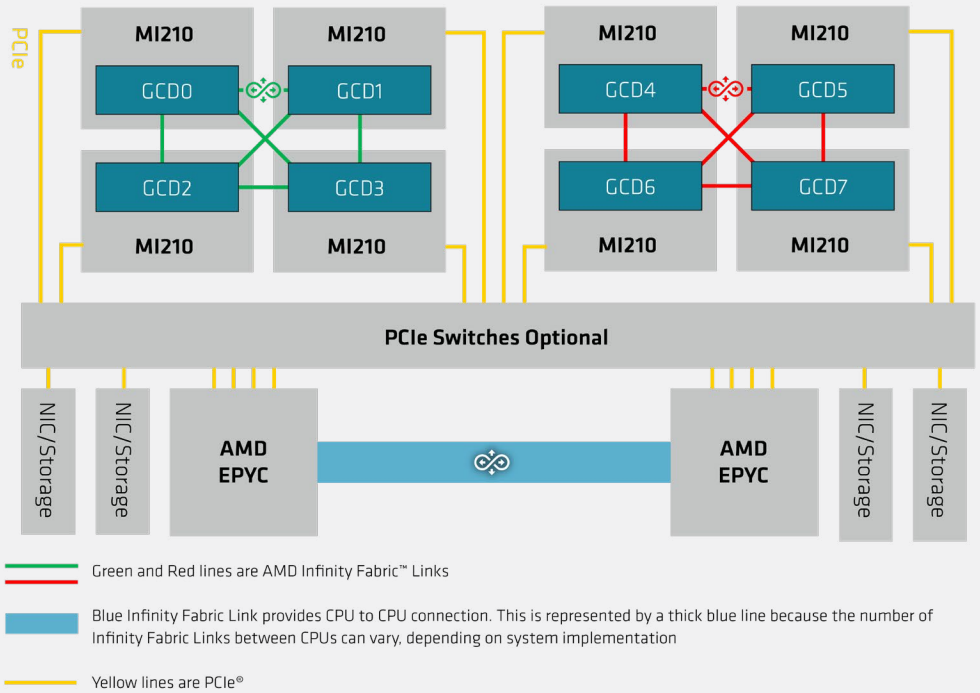


Figure 2e - Block diagram of a flagship HPC/AI server node built using the AMD Instinct™ MI210 accelerators in a quad GPU hive with direct connect xGMI bridge boards and AMD EPYC™ processor

AMD CDNA™ 2 Architecture Shader Array

Like AMD CDNA, in AMD CDNA™ 2, the command processor receives API commands and transforms them into compute tasks. The compute tasks are managed by the four Asynchronous Compute Engines (ACEs), which dispatch compute shader wavefronts to the compute units.

The AMD CDNA 2 compute units generally take an evolutionary approach and build on the strong foundation of prior generations as illustrated in Figure 3. The original AMD CDNA architecture was derived from the earlier GCN architecture and introduced the notion of matrices as a first-class citizen by adding the Matrix Core technology and support for new data types. AMD CDNA 2 doubles down on this approach and enhances several other aspects.

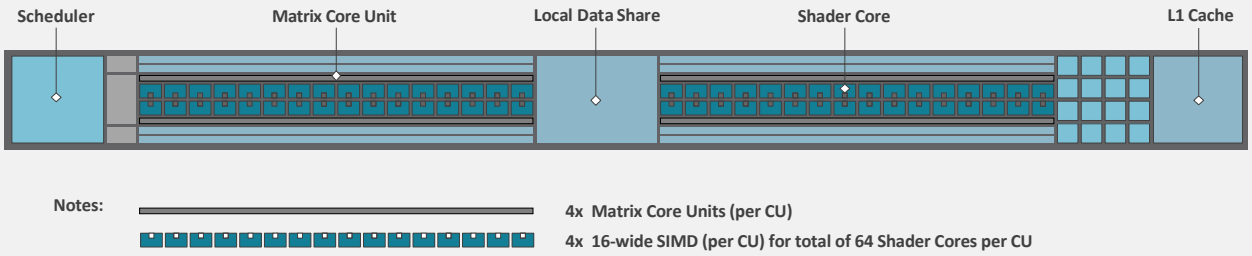


Figure 3 – Conceptional Block Diagram of an Enhanced Compute Unit (CU) with SIMD view of the AMD CDNA™ 2 architecture

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).

The AMD CDNA 2 compute units have been optimized carefully to boost the performance for scientific computing and provide more consistent throughput for datatypes used in machine learning. Historically, GPUs have been optimized for single-precision floating-point operations, and double-precision operations run at lower speed, ranging from half speed to as slow as one sixteenth. To improve performance for scientific computing applications, the AMD CDNA 2 vector pipeline has been tuned so that operating on wider double-precision data is the same rate as single-precision, with 64 fused multiply-add (FMA) operations per clock. Taking advantage of this improvement, the vector pipelines can also execute operations on packed single-precision values, doubling throughput to 128 single precision FMA operations per cycle.

The local data share is designed for explicitly passing data within a compute unit. This communication flow creates a natural opportunity for very high-throughput distributed atomic operations. In the AMD CDNA 2 architecture, the atomic execution units in the LDS have been improved to increase the throughput for FP64 min, max, and add atomic operations, which are commonly used in scientific computing e.g., molecular dynamics.

Computation	MI100 (FLOPS/CLOCK/CU)	MI250X (FLOPS/CLOCK/CU)	MI100 (Peak)	MI250X (Peak)	MI200 Peak Speedup
MI200 Matrix FP64 vs. MI100 Vector FP64	64	256	11.5 TFLOPS	95.7 TFLOPS	8.3x
MI200 Vector FP64 vs. MI100 Vector FP64	64	128	11.5 TFLOPS	47.9 TFLOPS	4.2x
MI200 Matrix FP32 vs. MI100 Matrix FP32	256	256	46.1 TFLOPS	95.7 TFLOPS	2.1x
MI200 Packed FP32 vs. MI100 Vector FP32	128	256	23.1 TFLOPS	95.7 TFLOPS	4.2x
MI200 Vector FP32 vs. MI100 Vector FP32	128	128	23.1 TFLOPS	47.9 TFLOPS	2.1x
MI200 Matrix FP16 vs. MI100 Matrix FP16	1024	1024	184.6 TFLOPS	383 TFLOPS	2.1x
MI200 Matrix BF16 vs. MI100 Matrix BF16	512	1024	92.3 TFLOPS	383 TFLOPS	4.2x
MI200 Matrix INT8 vs. MI100 Matrix INT8	1024	1024	184.6 TOPS	383 TOPS	2.1x

Table 1 – Generational comparison of numerical formats and peak throughput between MI250X (OAM) and MI100 (PCIe).

AMD CDNA™ 2 Matrix Core Technology

The AMD CDNA™ 2 architecture Matrix Core technology has also been enhanced, with an emphasis on high-performance computing. The AMD CDNA 2 Matrix Core technology now supports double-precision data, which is critical for many scientific computing applications. Matrix-matrix multiplication is one of the important primitives that can be leveraged in HPC kernels. Its accelerated implementation can speed up execution of HPC applications, including the important High Performance Linpack (HPL). Performing matrix multiplication using general FMA64 instructions is less efficient, spending substantial energy on register file accesses for each operand. Ultimately, this energy use limits the maximum performance that is possible within a given TDP.

AMD CDNA 2 introduces a set of matrix multiplication instructions specifically for FP64 precision with a simplified microarchitecture. New instructions realize block-based matrix multiplication for the fixed matrix blocks sizes of 16x16x4 and 4x4x4 (MxNxK) and are wave-wide operations where input and output matrix block data are distributed over a wavefront's lanes. The whole input is read from registers once and reused several times during calculation for a substantial reduction in power.

The FP64 matrix multiply instructions can deliver two times the throughput compared to using FP64 vector instructions, while also helping improve power efficiency. The net result is a corresponding 4X improvement in FP64 TFLOP/s compared to the MI100. These instructions can be used in AMD-provided libraries to accelerate linear algebra calculations and are expected to demonstrate maximal FP64 throughput for MI200. As Table 2 above illustrates, this quadruples the computational throughput for FP64 compared to the previous generation.

Additionally, the AMD CDNA 2 Matrix Core technology has improved the bfloat16 performance so that it delivers equivalent throughput to FP16.

AMD CDNA™ 2 Packed FP32

Another first in the AMD CDNA™ 2 architecture is Packed FP32, which executes two component vector instructions on FP32 operands for FMA, FADD and FMUL operations. These new instructions double vector FP32 throughput per clock per CU for those operations. The packed FP32 instructions rely on the vector operands being adjacent and aligned to even registers and apply the same rounding and denorm modes to both operations. To facilitate placing scattered scalar operands together into such vector, the architecture supports a dedicated packed move instruction that accesses two scalar registers and copies them to a pair of adjacent location registers. The result of this move operation can become an input of a packed math operation in a relatively straightforward way. The sample code below shows the modifications needed to take advantage of Packed FP32.

Original

```
float vxi = 0.0f, vyi = 0.0f, vzi = 0.0f;
for (int j = threadIdx.x; j < count1; j += blockDim.x) {
  float dx = xx1[j] - xxi;
  float dy = yy1[j] - yyi;
  float dz = zz1[j] - zzi;
  float dist2 = dx*dx + dy*dy + dz*dz;
  if (dist2 < fsrrmax2) {
    float rtemp = (dist2 + rsm2)*(dist2 + rsm2)*(dist2 + rsm2);
    float f_over_r = massi*mass1[j]*(1.0f/sqrt(rtemp) –
        (ma0 + dist2*(ma1 + dist2*(ma2 + dist2*(ma3 +
        dist2*(ma4 + dist2*ma5))))));
    vxi += fcoeff*f_over_r*dx;
    vyi += fcoeff*f_over_r*dy;
    vzi += fcoeff*f_over_r*dz;
  }
}
```

Modified to use Packed FMA32

```
float vxi = 0.0f, vyi = 0.0f, vzi = 0.0f;
  for (int j = threadIdx.x; j < count1; j += 2*blockDim.x) {
    float2 dx = {xx1[j] - xxi, xx1[j + blockDim.x] - xxi};
    float2 dy = {yy1[j] - yyi, yy1[j + blockDim.x] - yyi};
    float2 dz = {zz1[j] - zzi, zz1[j + blockDim.x] - zzi};
    float2 dist2 = dx*dx + dy*dy + dz*dz;
    bool check[2] = {dist2.x < fsrrmax2, dist2.y < fsrrmax2};
    if (check[0] || check[1]) {
      float2 rtemp = (dist2 + rsm2)*(dist2 + rsm2)*(dist2 + rsm2);
      float2 mass1_2 = {mass1[j],mass1[j + blockDim.x]};
      float2 sqrt_rtemp = {sqrtf(rtemp.x),sqrtf(rtemp.y)};
      float2 f_over_r = massi*mass1_2*(1.0f/sqrt_rtemp –
          (ma0 + dist2*(ma1 + dist2*(ma2 + dist2*(ma3 +
          dist2*(ma4 + dist2*ma5))))));
      float2 vxi_tmp = fcoeff*f_over_r*dx;
      float2 vyi_tmp = fcoeff*f_over_r*dy;
      float2 vzi_tmp = fcoeff*f_over_r*dz;
      vxi += check[0] ? vxi_tmp.x : 0.0f;
      vxi += check[1] ? vxi_tmp.y : 0.0f;
      vyi += check[0] ? vyi_tmp.x : 0.0f;
      vyi += check[1] ? vyi_tmp.y : 0.0f;
      vzi += check[0] ? vzi_tmp.x : 0.0f;
      vzi += check[1] ? vzi_tmp.y : 0.0f;
    }
  }
```

Figure 4 –Sample code showing the modifications needed to take advantage of Packed FP32.

AMD ROCm™ Open Software Platform Enables AMD CDNA™ 2

The key to accelerated computing for HPC and ML is a software stack and ecosystem that easily unlocks the capabilities for software developers and customers. AMD ROCm™ stack, shown in Figure 5, provides an open-source and easy to use set of tools that are built around industry standards and enable creating well-optimized portable software for everything from simple workstation programs to massive exascale applications.

The principles behind AMD ROCm are fairly simple. First, accelerated computing requires equality between both processors and accelerators. While they focus on different workloads, they should work together effectively and have equal access to resources such as memory. Second, a rich ecosystem of software libraries and tools should enable portable and performant code that can take advantage of new capabilities. Last, an open-source approach empowers vendors, customers, and the entire community along with amplification of AMD's own investment.

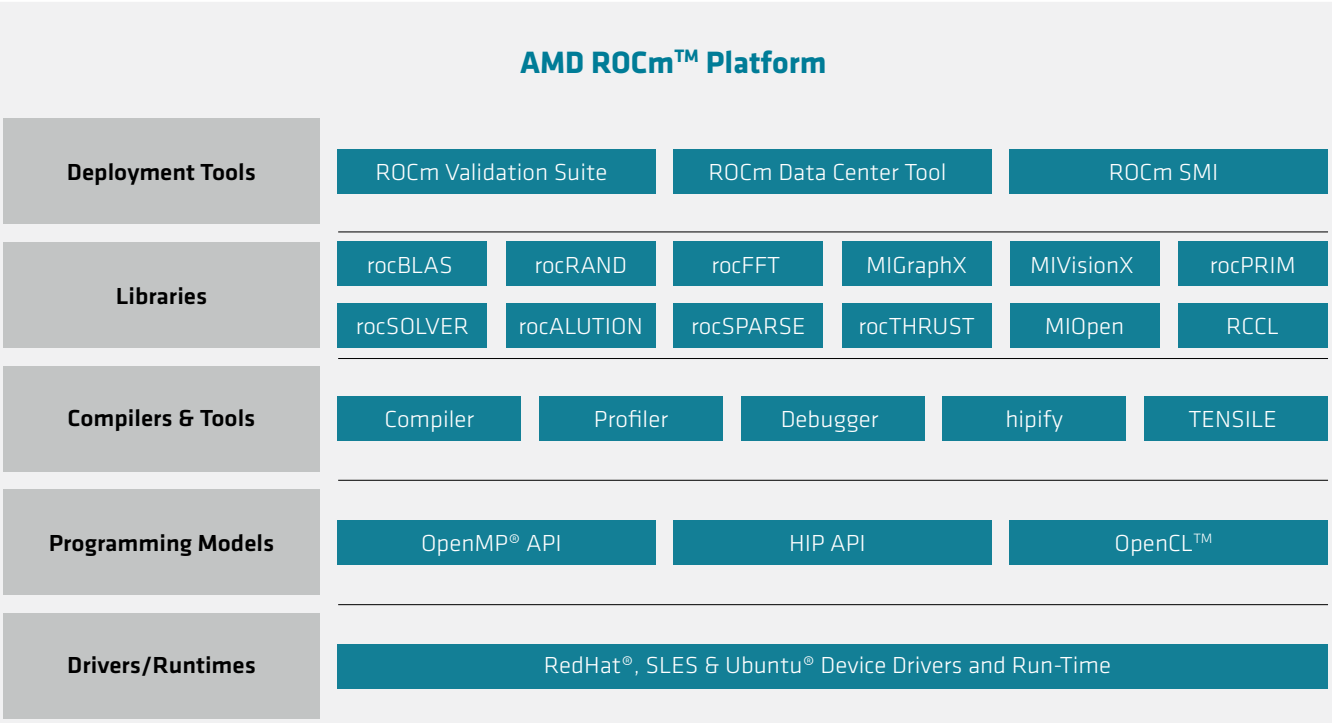
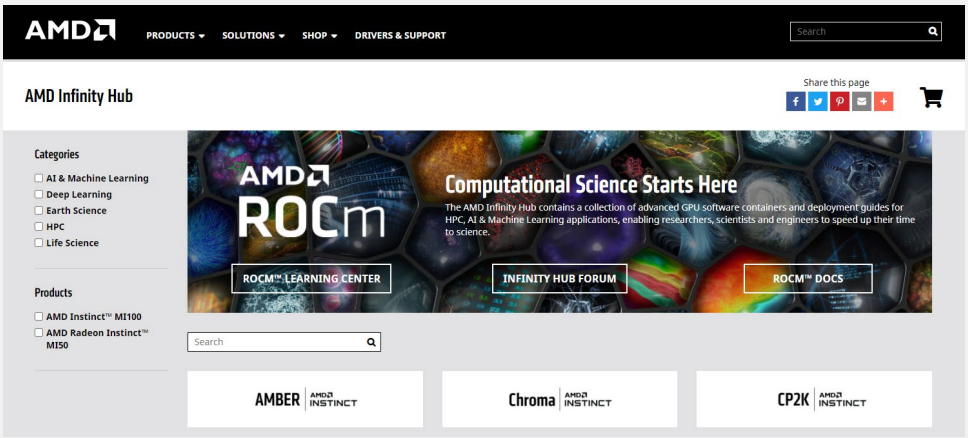


Figure 5 – AMD's open-source ROCm stack includes tools developers need to build high-performance applications for scientific computing and machine learning.

The AMD ROCm™ ecosystem is crucial for putting the capabilities of the AMD CDNA™ 2 architecture into the hands of developers, vendors, customers, and the entire community. For example, the rocBLAS library has incorporated the new instructions for FP64 matrix-multiplication and packed FP32 vectors so that developers that work with higher-level libraries will get excellent performance from day one. At a lower level, the ROCm compiler and runtime can take advantage of these same features to generate high-performance binaries for custom code and a more diverse set of applications beyond linear algebra. At an even higher level, AMD's Infinity Hub (https://www.amd.com/en/technologies/infinity-hub) contains containerized HPC and ML applications that are ready to use and support the latest MI200 series accelerators.

At the same time, the unique capabilities of the AMD CDNA 2 architecture - especially cache coherency, enable simplifying applications and delivering even greater performance. For example, parts of NWChemEx make use of coherent unified memory; porting this to non-coherent processors and accelerators could add complexity, introduce new bugs, and generally delay deploying the application. AMD MI250X accelerator with the optimized 3rd gen AMD EPYC™ processor in a cache-coherent configuration can greatly improve productivity.



For other applications, the fine-grained communication boosts performance. For example, the HACC[6] cosmology simulator, builds a tree data structure to track particles and their nearest-neighbors. These relationships are subsequently used to compute N-body gravitational forces between the particles. While calculating the N-body forces is a perfect fit for an accelerator, building the tree is best done on a processor that can handle branchy code with challenging data locality. A cache coherent accelerator like the MI250X can read the particle tree to start spawning work while the optimized 3rd gen AMD EPYC™ processor is simultaneously building other portions of the tree, avoiding explicit synchronization and stalls, and reducing expensive copies of the entire tree. The fine-grained serialization enabled by cache coherency can boost utilization and performance compared to a non-coherent accelerator. Since cache coherent accelerators are relatively new, the AMD MI250X coupled with the optimized 3rd gen AMD EPYC processor will be a crucial platform for the industry to explore the possibilities and understand the benefits of cache coherency.

AMD Instinct MI200 Series Accelerator Product Offerings

Performance	MI210	MI250	MI250X
CDNA2 Graphics Compute Die (GCD)	1	2	2
Compute Units	104 CU	208CU	220CU
Stream processors	6,656	13,312	14,080
Matrix Cores	416	832	880
Peak FP64/FP32 Vector	22.6 TF	45.3 TF	47.9 TF
Peak FP64/FP32 Matrix	45.3 TF	90.5 TF	95.7 TF
Peak FP16/BF16	181.0 TF	362.1 TF	383.0 TF
Peak INT4/INT8	181.0 TOPS	362.1 TOPS	383.0 TOPS
Memory			
Memory Size	64GB HBM2e	128GB HBM2e	128GB HBM2e
Memory Interface	4,096 bits	8,192 bits	8,192 bits
Memory Clock	1.6GHz	1.6GHz	1.6GHz
Memory Bandwidth	up to 1.6 TB/sec	up to 3.2TB/sec2	up to 3.2TB/sec2
Scalability			
Infinity Fabric Links	up to 3	up to 6	up to 8
xGMI Bridge Card Configuration	Yes (Dual | Quad Hives)	NA	NA
Coherency Enabled	No	No	Yes
Reliability			
ECC (Full-chip)	Yes	Yes	Yes
RAS Support	Yes	Yes	Yes
Board Design			
Board Form Factor	PCIe Full-Height, Full-Length (Dual Slot)	OAM	OAM
Bus Interface	PCIe® Gen4 Support	PCIe® Gen4 Support	PCIe® Gen4 Support
Thermal	Passively Cooled	Passive & Liquid	Passive & Liquid
Max Power	300W TDP	500W & 560W TDP	500W & 560W TDP
Warranty	Three Year Limited	Three Year Limited	Three Year Limited

Table 1 -

Conclusion

The era of exascale computing will push the edges of human discovery in scientific computing and machine learning, unlocking new innovations that will benefit the world. Heterogeneous computing systems are undeniably the key to this leap forward in computational performance. GPUs have evolved from simple fixed-function hardware pipelines to fully programmable and general-purpose accelerators that are a critical element of any heterogeneous system.

AMD CDNA™ 2 architecture is a key step forward for accelerators and heterogeneous systems - one that will unlock the possibilities of exascale computing. The AMD CDNA 2 architecture significantly improves computational throughput, delivering nearly 48 TFLOP/s peak theoretical double-precision compute in a single accelerator[1] . This up to 4X generational improvement in compute[1] is achieved through refinements to the compute units and more radical innovations in communication that enable nearly doubling performance through AMD's Infinity Fabric™ technology and multi-chip packaging and boosting full-node scalability.

These same radical improvements allow accelerators such as the AMD Instinct™ MI250X to not only share memory with an optimized 3rd gen AMD EPYC™ processor, but to offer full cache coherency and act as a fully peer element of a heterogeneous system. This is a tremendous step forward in programmability that also enables unique optimizations and creates a testbed for the industry to take advantage of cache coherency across many different workloads. Best of all, the advantages of the AMD CDNA 2 architecture will be readily available to vendors, customers, and the entire community through the open-source AMD ROCm™ ecosystem.

AMD Resources

AMD Instinct™ MI200 series accelerators: www.amd.com/Instinct

AMD CDNA™ 2 Architecture: www.amd.com/en/technologies/cdna2

AMD Infinity Architecture: www.amd.com/en/technologies/Infinity-Architecture

AMD ROCm™ open software platform: www.amd.com/ROCm

AMD Infinity Hub: www.amd.com/en/technologies/infinity-hub

ROCm™ Learning Center: www.developer.amd.com/resources/rocm-learning-center/

Acronyms:

ACE - Asynchronous Compute Engine
AI - Artificial Intelligence
AVC - Advanced Video Coding
CPU - Central Processing Unit
ESM - Extended Speed Mode
FMA - Fused Multiply-Add
GCD - Graphics Compute Die
GPU - Graphics Processing Unit
HEVC - High Efficiency Video Coding
HPC - High Performance Computing
LDS - Local Data Store
ML- Machine Learning
OAM - Open Compute Project Accelerator Module
PCIe - PCI-Express
SIMD - Single Instruction, Multiple Data
TDP - Thermal Design Power
TFLOPS - Trillions Floating Point Operations per Second
TOPS - Trillions Operations per Second
VCN - Video Codec Next

Endnotes:

1. Measurements conducted by AMD Performance Labs as of Sep 10, 2021 on the AMD Instinct™ MI250X accelerator designed with AMD CDNA™ 2 6nm FinFET process technology with 1,700 MHz engine clock resulted in 47.9 TFLOPS peak double precision (FP64) floating-point, 383.0 TFLOPS peak Bfloat16 format (BF16) floating-point performance. The results calculated for AMD Instinct™ MI100 GPU designed with AMD CDNA 7nm FinFET process technology with 1,502 MHz engine clock resulted in 11.54 TFLOPS peak double precision (FP64) floating-point, 92.28 TFLOPS peak Bfloat16 format (BF16) performance. MI200-05
2. The AMD Instinct™ MI250X accelerator has 220 compute units (CUs) and 14,080 stream cores. The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores. MI200-27
3. Video codec acceleration (including at least the HEVC (H.265), H.264, VP9, and AV1 codecs) is subject to and not operable without inclusion/installation of compatible media players. GD-176
4. Calculations conducted by AMD Performance Labs as of Oct 29th, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e OAM Module) 500W and 560W accelerators at 1,700 MHz peak boost engine clock designed with AMD CDNA™ 2 6nm FinFet process technology resulted in 6.96 TB/s peak theoretical L2 cache slice bandwidth performance. Calculations by AMD Performance Labs as of OCT 5th, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe®) 300W accelerator at 1,502 MHz peak boost engine clock accelerator designed with AMD CDNA 7nm FinFET process technology resulted in 3.07 TB/s peak theoretical L2 cache slice bandwidth performance. MI200-34
5. Calculations conducted by AMD Performance Labs as of Oct 18th, 2021, for the AMD Instinct™ MI250X and MI250 accelerators (OAM) designed with CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 128GB HBM2e memory capacity and 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250X/MI250 memory bus interface is 8,192 bits and memory data rate is up to 3.20 Gbps for total memory bandwidth of 3.2768 TB/s. Calculations by AMD Performance Labs as of OCT 18th, 2021 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 32GB HBM2 memory capacity and 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI100 memory bus interface is 4,096 bits and memory data rate is up to 2.40 Gbps for total memory bandwidth of 1.2288 TB/s.MI200-30
6. GROMACS: http://www.gromacs.org/
7. HACC: https://cpac.hep.anl.gov/projects/hacc/
8. Calculations as of Oct 18th, 2021. AMD Instinct™ MI250/MI250X built on AMD CDNA™ 2 technology accelerators support AMD Infinity architecture with AMD Infinity Fabric™ technology providing up to 400 GB/s total aggregate theoretical inter GDC to GDC I/O data transport bandwidth per GPU. Peak theoretical inter GDC to GDC data transport rate performance is calculated by Baud Rate * # lanes * # directions * # links / 8 = GB/s per card. MI200-29
9. Calculations as of Sep 18th, 2021. AMD Instinct™ MI250 built on AMD CDNA™ 2 technology accelerators support AMD Infinity Fabric™ technology providing up to 100 GB/s peak total aggregate theoretical transport data GPU peer-to-peer (P2P) bandwidth per AMD Infinity Fabric link, and include up to eight links providing up to 800GB/s peak aggregate theoretical GPU (P2P) transport rate bandwidth performance per GPU OAM card for 800 GB/s. AMD Instinct™ MI100 built on AMD CDNA technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card, and include three links providing up to 276 GB/s peak theoretical GPU P2P transport rate bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. Server manufacturers may vary configuration offerings yielding different results. MI200-13
10. Calculations conducted by AMD Performance Labs as of Jan 27, 2022, for the AMD Instinct™ MI210 (64GB HBM2e) accelerator (PCIe®) designed with AMD CDNA™ 2 architecture 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 64 GB HBM2e memory capacity and 1.6384 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 3.20 Gbps for total memory bandwidth of 1.6384 TB/s ((3.20 Gbps*(4,096 bits))/8). Calculations conducted by AMD Performance Labs as of Sep 18, 2020, for the AMD Instinct™ MI100 (32GB HBM2) accelerator (PCIe®) designed with AMD CDNA™ architecture 7nm FinFet process technology at 1,502 MHz peak clock resulted in 32 GB HBM2 memory capacity and 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 2.40 Gbps for total memory bandwidth of 1.2288 TB/s ((2.40 Gbps*(4,096 bits))/8). MI200-42

11. Calculations as of JAN 27th, 2022. AMD Instinct™ MI210 built on AMD CDNA™ 2 technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical data bandwidth from CPU to GPU per card. AMD Instinct™ MI210 CDNA 2 technology-based accelerators include three Infinity Fabric™ links providing up to 300 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 364 GB/s. Dual-GPU hives: One dual-GPU hive provides up to 300 GB/s peak theoretical P2P performance. Four-GPU hives: One four-GPU hive provide up to 600 GB/s peak theoretical P2P performance. Dual four GPU hives in a server provide up to 1.2 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provide up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0. AMD Instinct™ MI100 built on AMD CDNA technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. AMD Instinct™ MI100 CDNA technology-based accelerators include three Infinity Fabric™ links providing up to 276 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. One four-GPU hive provides up to 552 GB/s peak theoretical P2P performance. Dual four-GPU hives in a server provide up to 1.1 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provides up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0. Server manufacturers may vary configuration offerings yielding different results. MI200-43