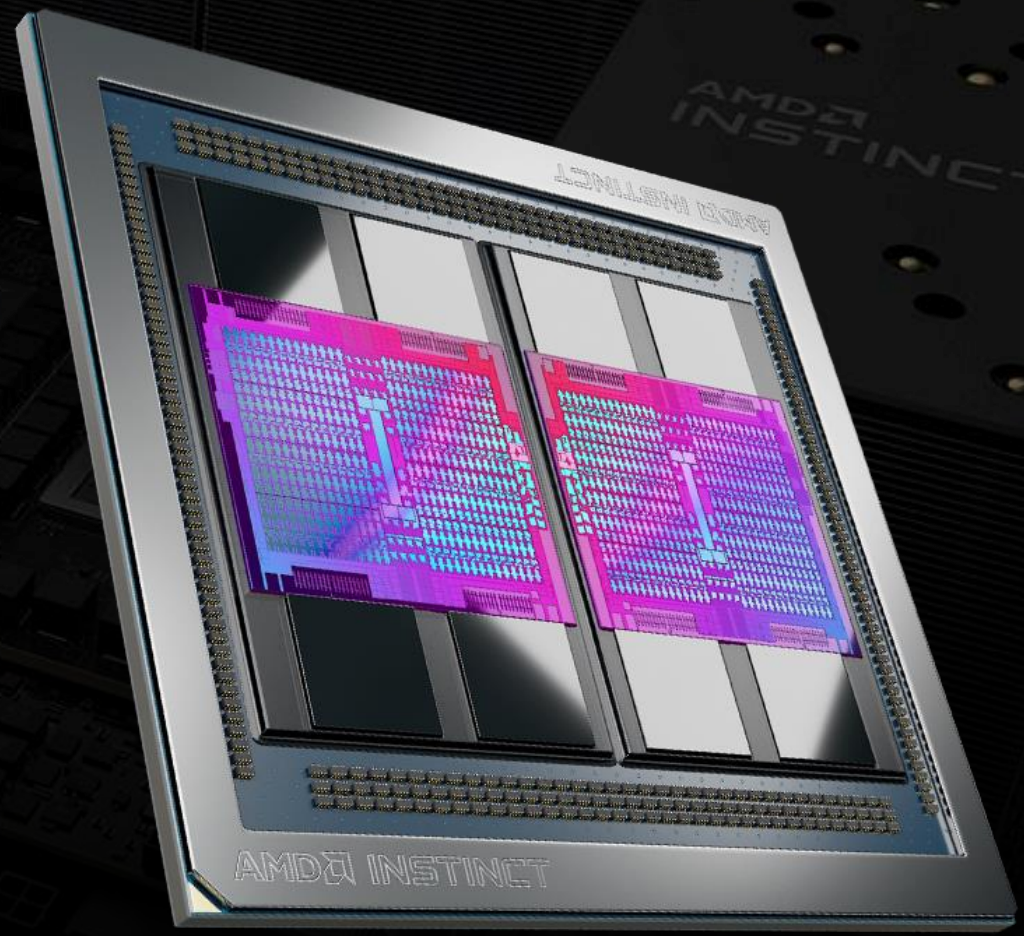
AMD Instinct™ MI200 Series Accelerator and Node Architectures

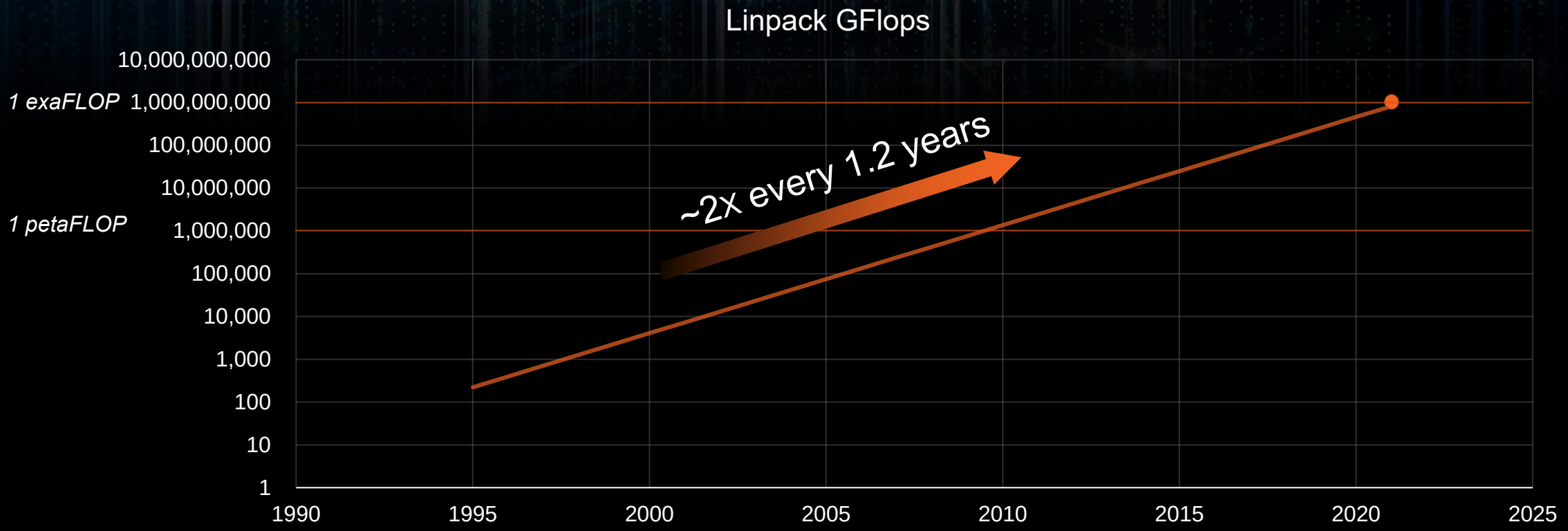
Alan Smith
AMD Sr. Fellow and Instinct Lead SOC Architect

Norman James
AMD Fellow and Instinct Lead System Architect

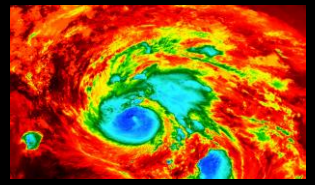


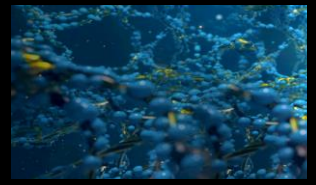
Relentless Demand for Scientific Computing

World's Fastest Supercomputers

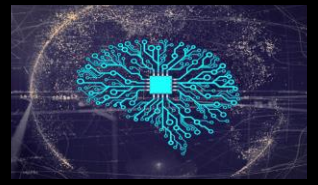


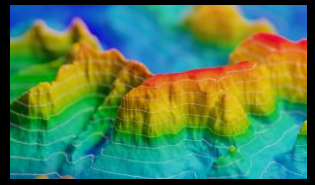

Space Exploration


Climate Change


Chemical Sciences


Energy Solutions

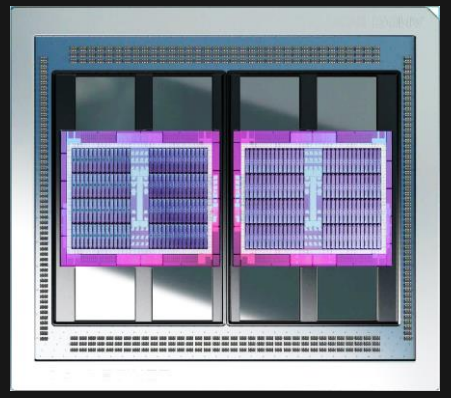

Machine Learning


Real Time Simulation

AMD

AMD Instinct™ MI200 Series

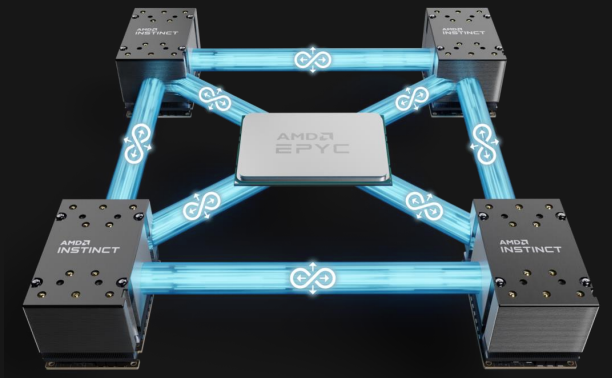
First Multi-die GPU



Workload-optimized Compute Architecture

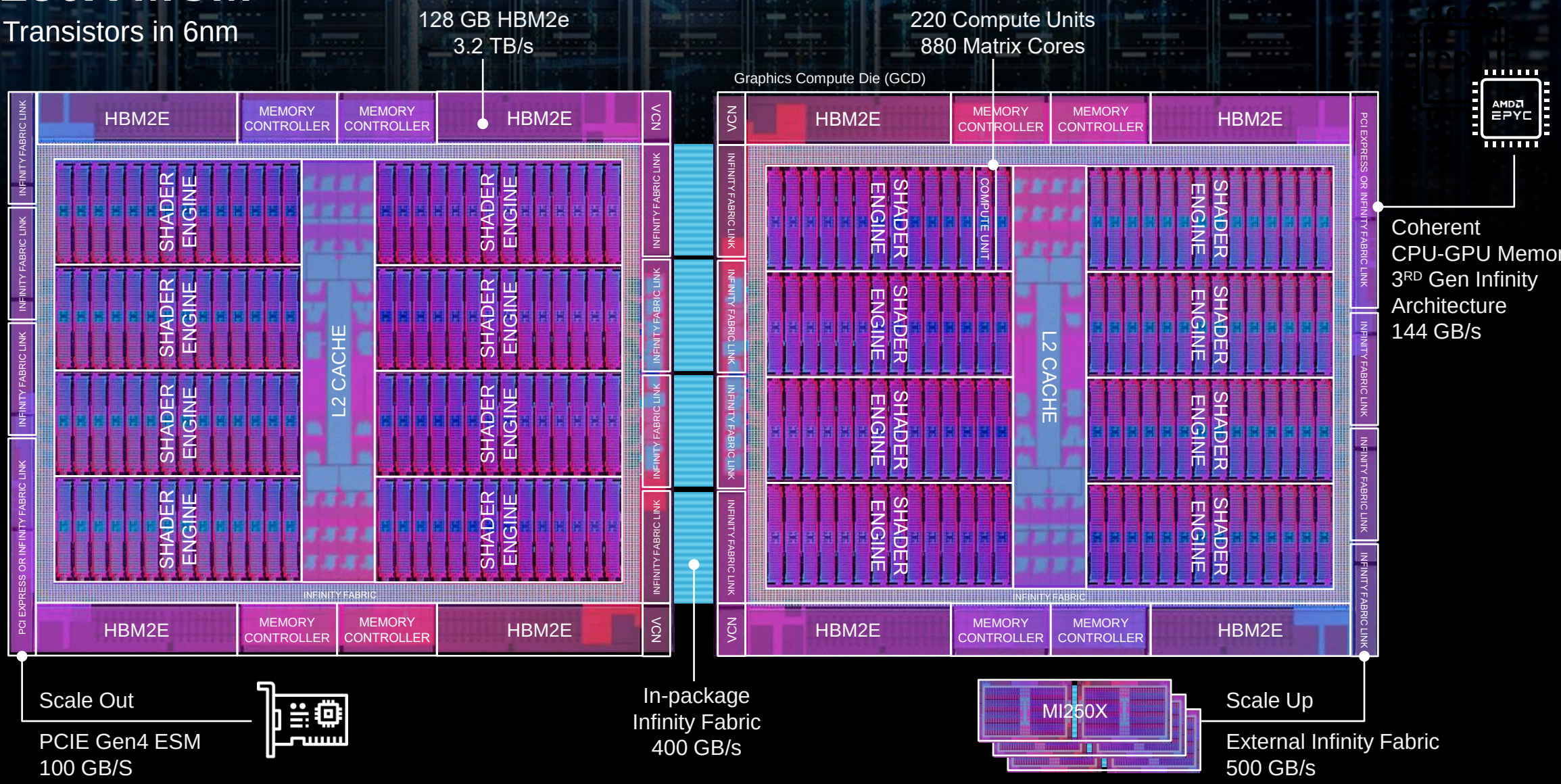
AMD CDNA 2

3rd Gen AMD Infinity Architecture



MI250X MCM

58B Transistors in 6nm



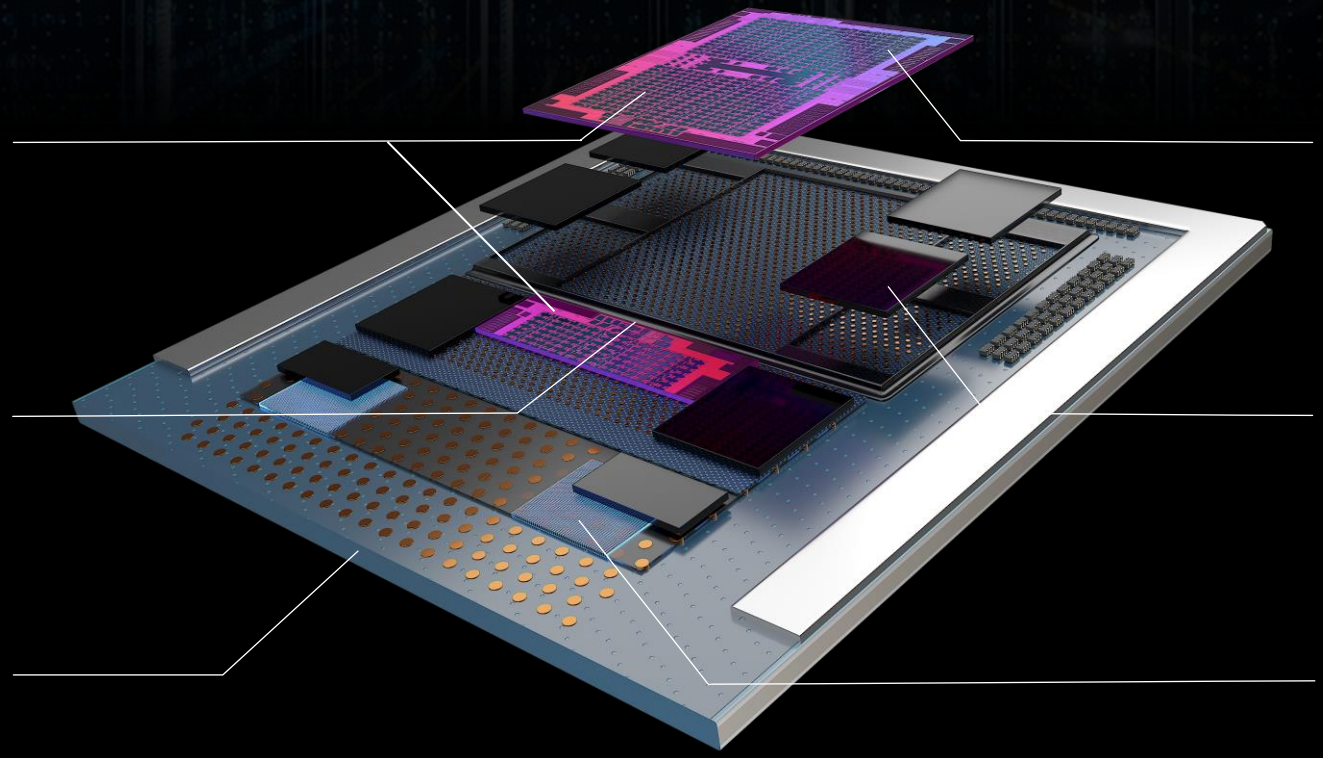
AMD Instinct™ MI200 Series

Key Innovations

Two
AMD CDNA™2 Dies

Ultra High Bandwidth
Die Interconnect

Coherent CPU-to-GPU
Interconnect



2nd Gen Matrix Cores
for HPC & AI

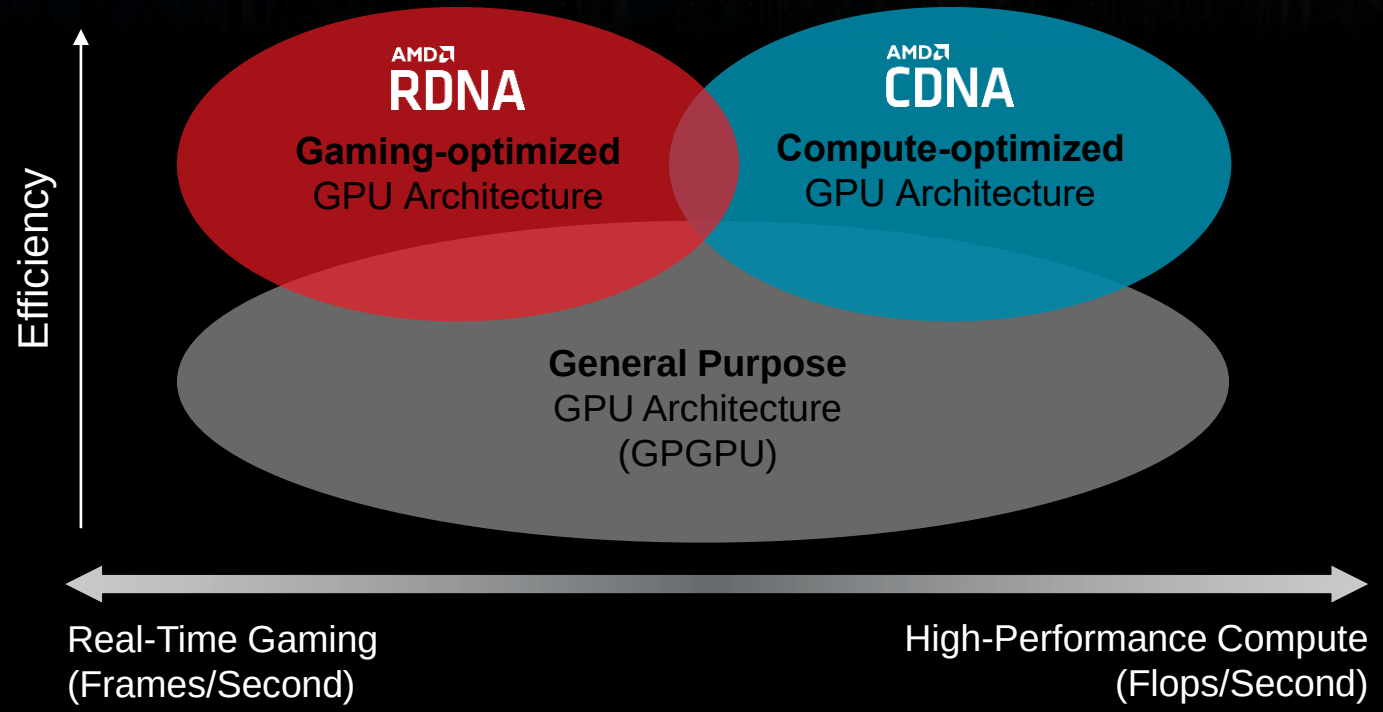
Eight Stacks
of HBM2E

2.5D Elevated
Fanout Bridge (EFB)

AMD Instinct™ MI200 OAM Series

Domain-specific Architectures

Optimal Efficiency through Domain-specific Optimization

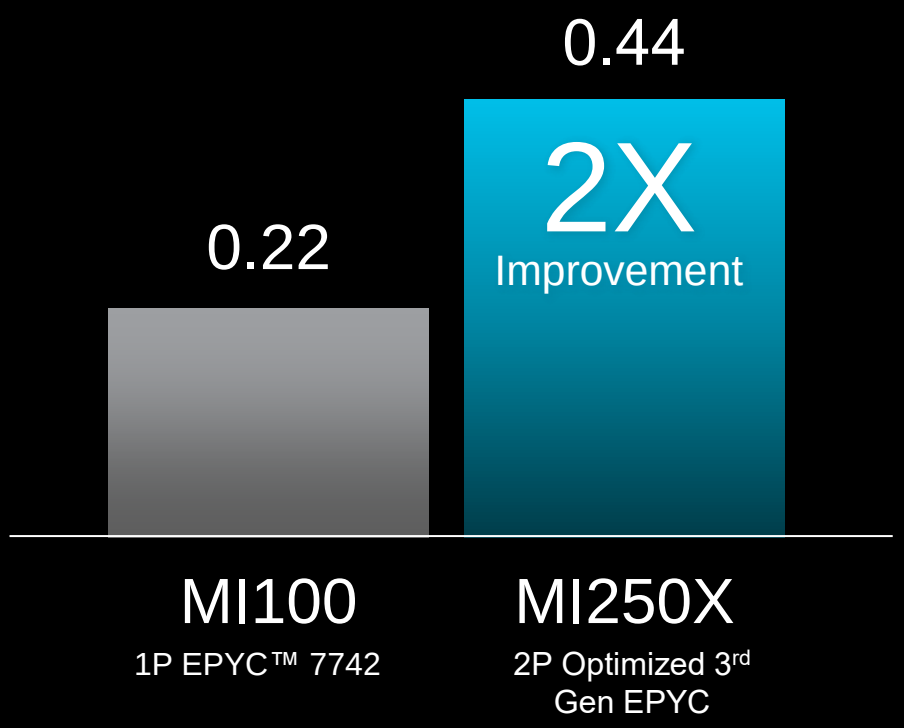


Moore's Law Slowing
Costs and Power Increasing

Optimal Efficiency
Architectures optimized for target workloads

AMD CDNA 2™ CU Performance-per-watt

Delivered Double Precision GEMM Performance (GFLOPS/WATT/CU)



Performance Contributors

Design Frequency Increase

- Leveraged CPU expertise
- Streamlined micro-architecture and design

Power Optimizations

- Tuned for low voltage operation
- Minimized clock and data movement power

Architecture Innovations

- Efficient matrix data-paths
- Extensive operand reuse and forwarding

See Endnote MI200-64

CDNA 2 Compute Unit with Enhanced Matrix Cores

Combined register resources with a custom SRAM design for matrix and vector ops

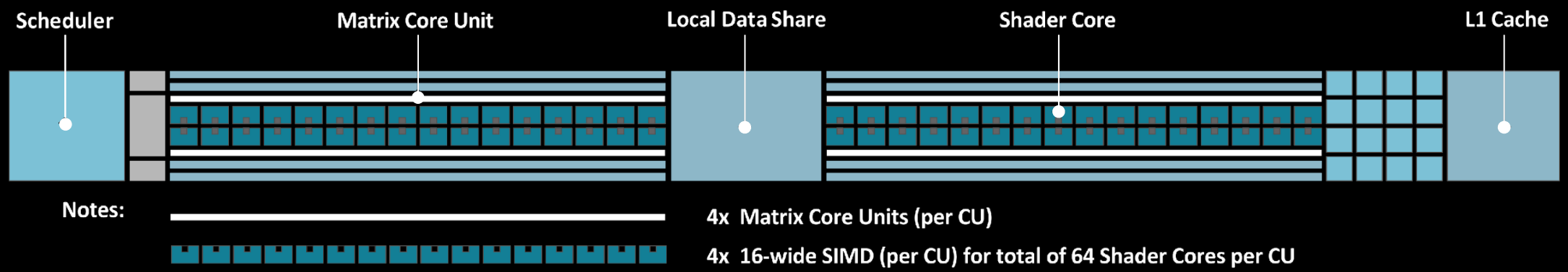
- Reduced energy on register file accesses and increased capacity for all operation types

Enhanced matrix cores; with emphasis on high-performance computing

- 4x double-precision and 2x Bfloat16 matrix OPS throughput relative to prior generation
- New power efficient IEEE754 compliant double-precision matrix instructions for 16x16x4 and 4x4x4 (MxNxK) blocks
- Full input operand reuse and output accumulator forwarding for substantial reduction in power

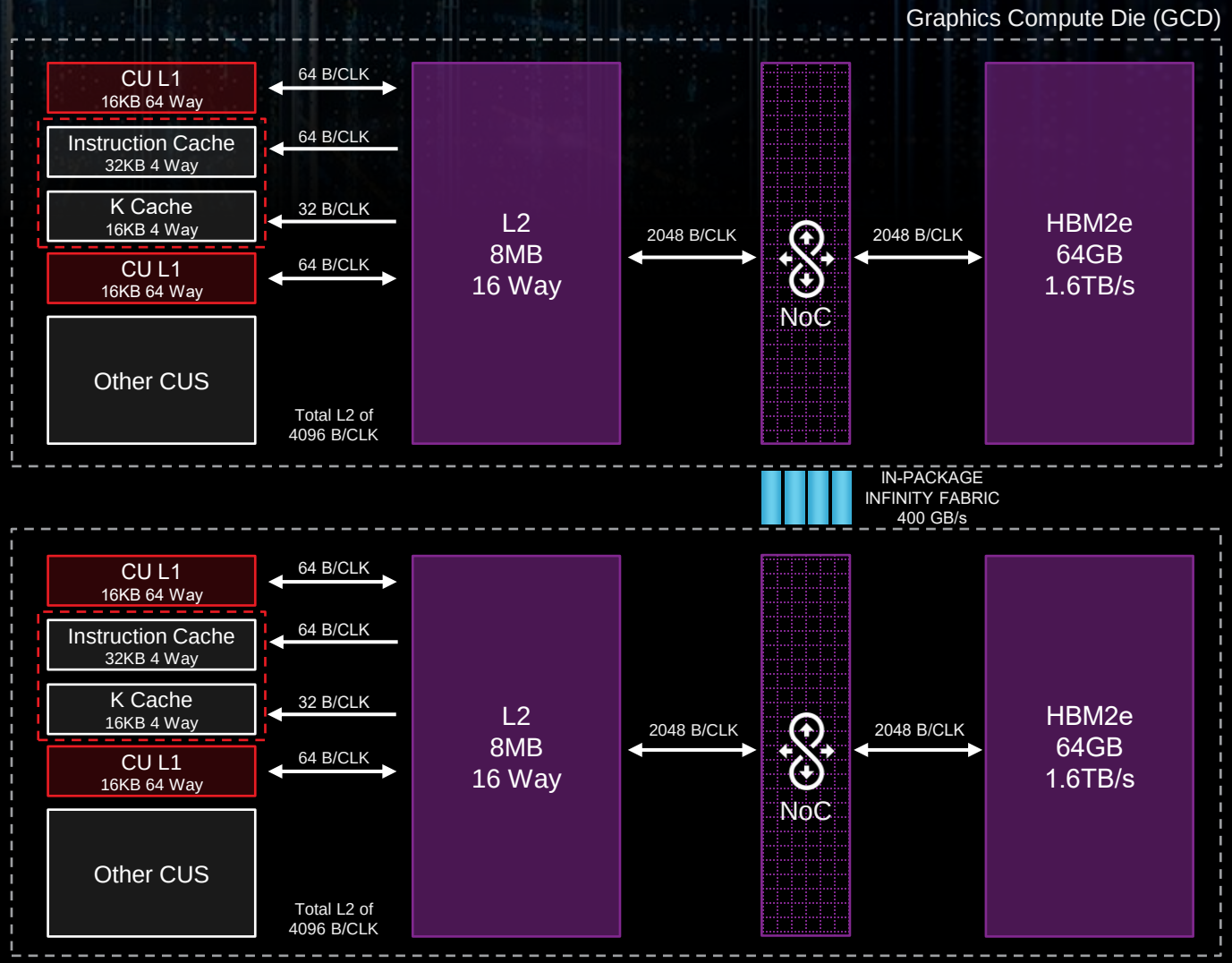
2x double-precision throughput and packed single-precision vector OPS relative to prior generation

- New instructions for packed single-precision vector operations (FMA, MUL, ADD, MOVE)



CDNA 2 Memory and Cache Hierarchy

- Scaling for throughput and large datasets
- Per GCD L2 Cache
  - 8MB total capacity
  - Physically partitioned into 32 slices
  - Each slice delivers 128B/CLK
  - Enhanced queuing and arbitration
  - Enhanced atomic operations
- Per GCD Memory Subsystem
  - 64GB HBM2e per GCD
  - Aggregate 1.6TB/s per GCD
  - Physically partitioned into 32 channels
  - 64B/CLK for efficient operating voltage
- In-package Infinity Fabric
  - Unified Shared Memory across GCDs
  - 400GB/s bisection bandwidth



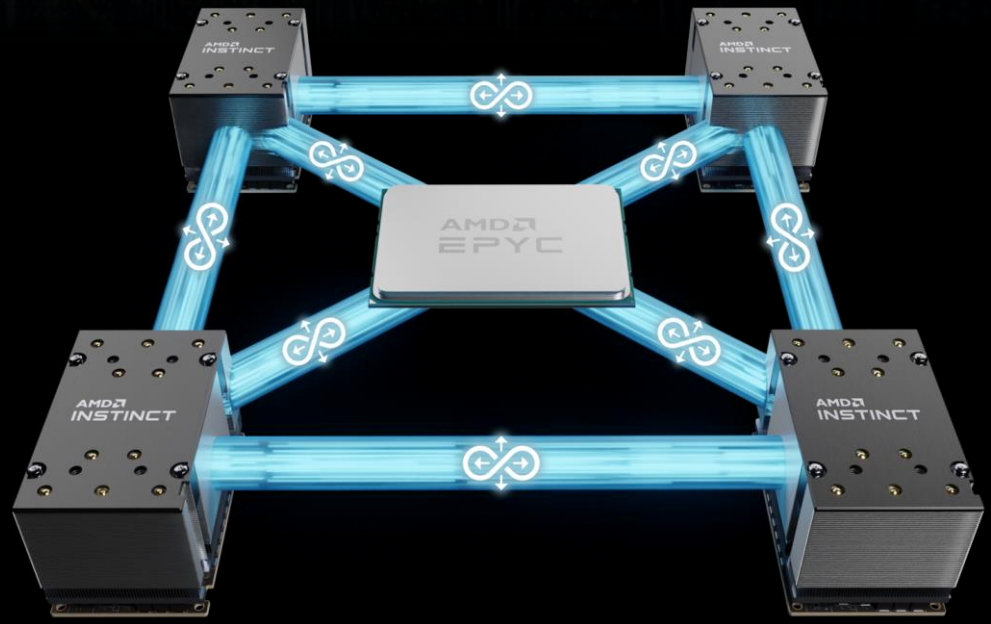
Shattering Performance Barriers in HPC & AI

Peak Performance	MI100* (Peak)	MI250X* (Peak)	MI250X Peak Speedup
FP64 Vector	11.5 TF	47.9 TF	4.2x
FP32 Vector	23.1 TF	47.9 TF	2.1x
Packed FP32 Vector	23.1 TF	95.7 TF	4.2x
FP64 Matrix	11.5 TF	95.7 TF	8.3x
FP32 Matrix	46.1 TF	95.7 TF	2.1x
BF16 Matrix	92.3 TF	383 TF	4.2x
FP16 Matrix	184.6 TF	383 TF	2.1x
Memory Size	32 GB	128 GB	4x
Memory Bandwidth	1.2 TB/s	3.2 TB/s	2.7x

* See Endnotes

3rd Generation AMD Infinity Architecture

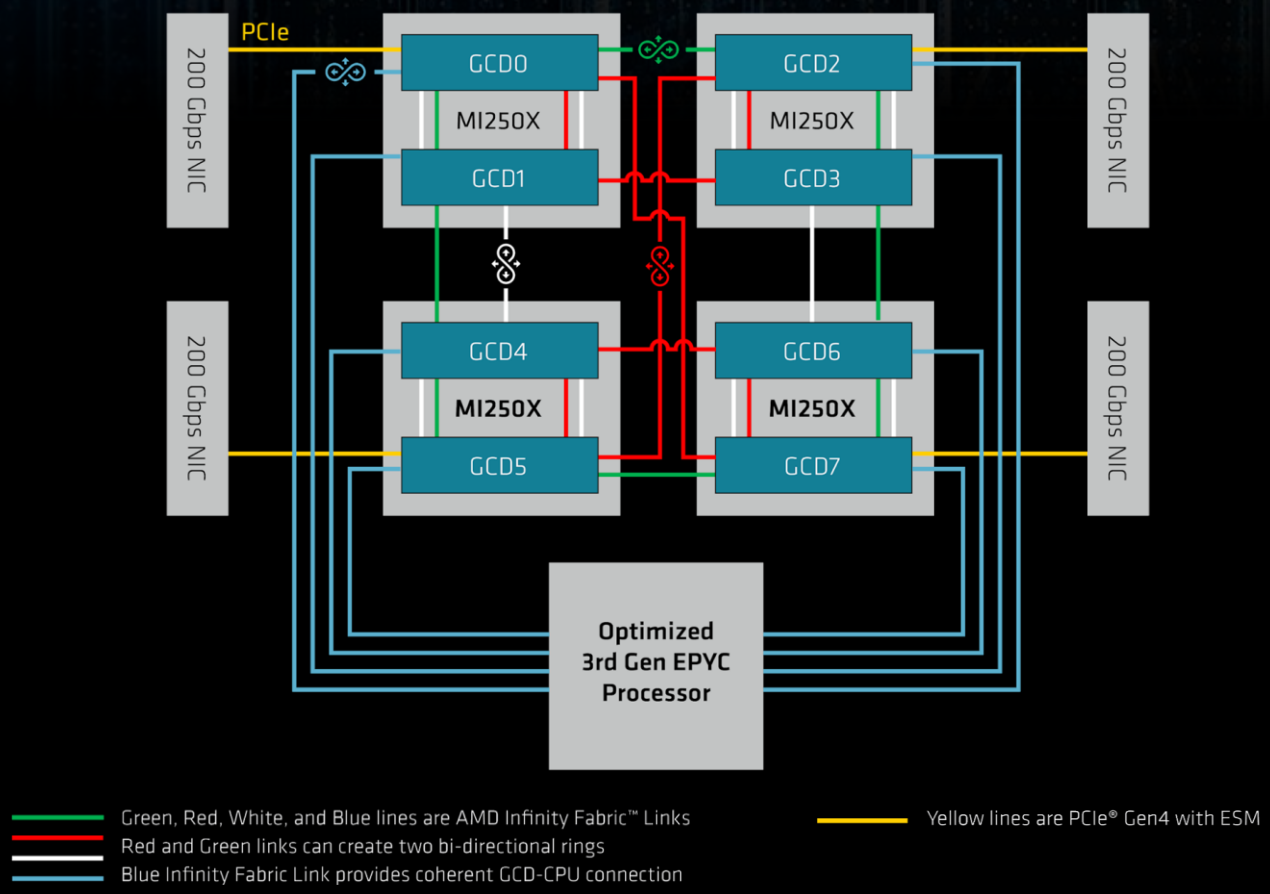
- Unified Shared Memory
- CPUs hardware coherently cache DDR and HBM memory
- GPUs hardware coherently cache local HBM memory
- Each GPU connects to host over 16b link at 18GT
  - Host memory BW is 200 GB/s (8 Ch DDR3200)
  - Each MI250X capable of saturating 2 Dram channels
- Allows NICs to be attached to MI250X
  - Enables line rate for PCIe ordered traffic to host memory, local HBM memory and peer HBM memory on the same socket
- Flat ID based routing across 8 GCDs and the host CPU
- Single large system with Root Complex on the device
- TLB shoot downs invalidates host and IOMMU page table entries across the entire system



MI200 SERIES NODE TOPOLOGIES AND SYSTEMS

Flagship HPC Topology with AMD Instinct™ MI250X GPU

A Unified Computing Node



One Optimized 3rd Gen
AMD EPYC™ Processor

Four AMD Instinct
MI250X Accelerators

5 GPU-to-GPU
Infinity Fabric Links

2 Coherent
CPU-to-GPU Links Per OAM

1 GPU Connected
PCIe® NIC Per OAM

1.54 TB/s
Peak Infinity Fabric Bandwidth

HPE CRAY EX235A Accelerator Blade

With AMD Instinct™ MI250X Accelerators and Optimized 3rd Gen AMD EPYC™ CPUs

Hewlett Packard Enterprise

Two Nodes Per Blade

Four AMD Instinct MI250X Per Node

One Optimized 3rd Gen AMD EPYC Per Node

AMD

MI250X Hits 1.1 ExaFlops!

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
★ 1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
★ 3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942

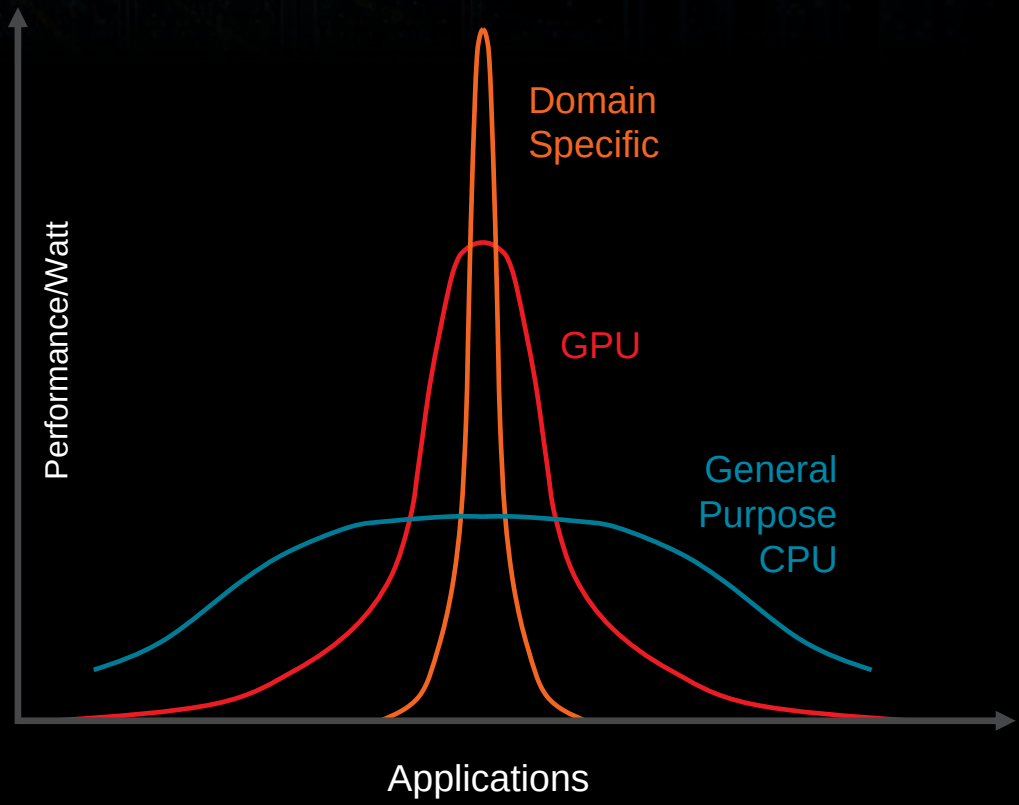
Source: TOP500 list, as of May 30,2022

Powerful Yet Efficient

MI250X Captures Top 4 Spots on the Green500 List

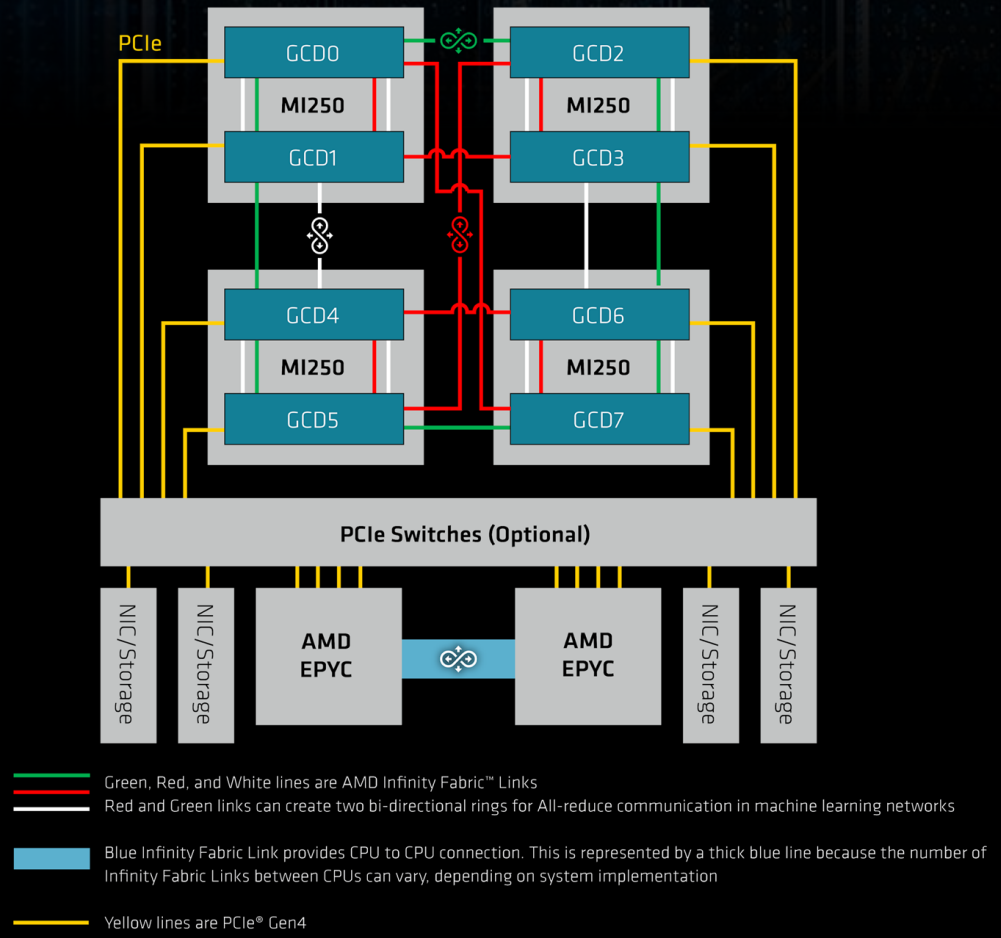
Rank	TOP500 Rank	System	Cores	Rmax (PFlop/s)	Power (kW)	Energy Efficiency (GFlops/watts)
1	29	**Frontier** TDS - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	120,832	19.20	309	62.684
2	1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	21,100	52.227
3	3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	2,942	51.629
4	10	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l'Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	921	50.028

Application-Specific Optimization Provides Better Performance-per-watt



Source: Green500 list, as of May 30,2022

Mainstream Accelerated Topology with AMD Instinct™ MI250 GPUs



Two 3rd Gen
AMD EPYC™ Processors

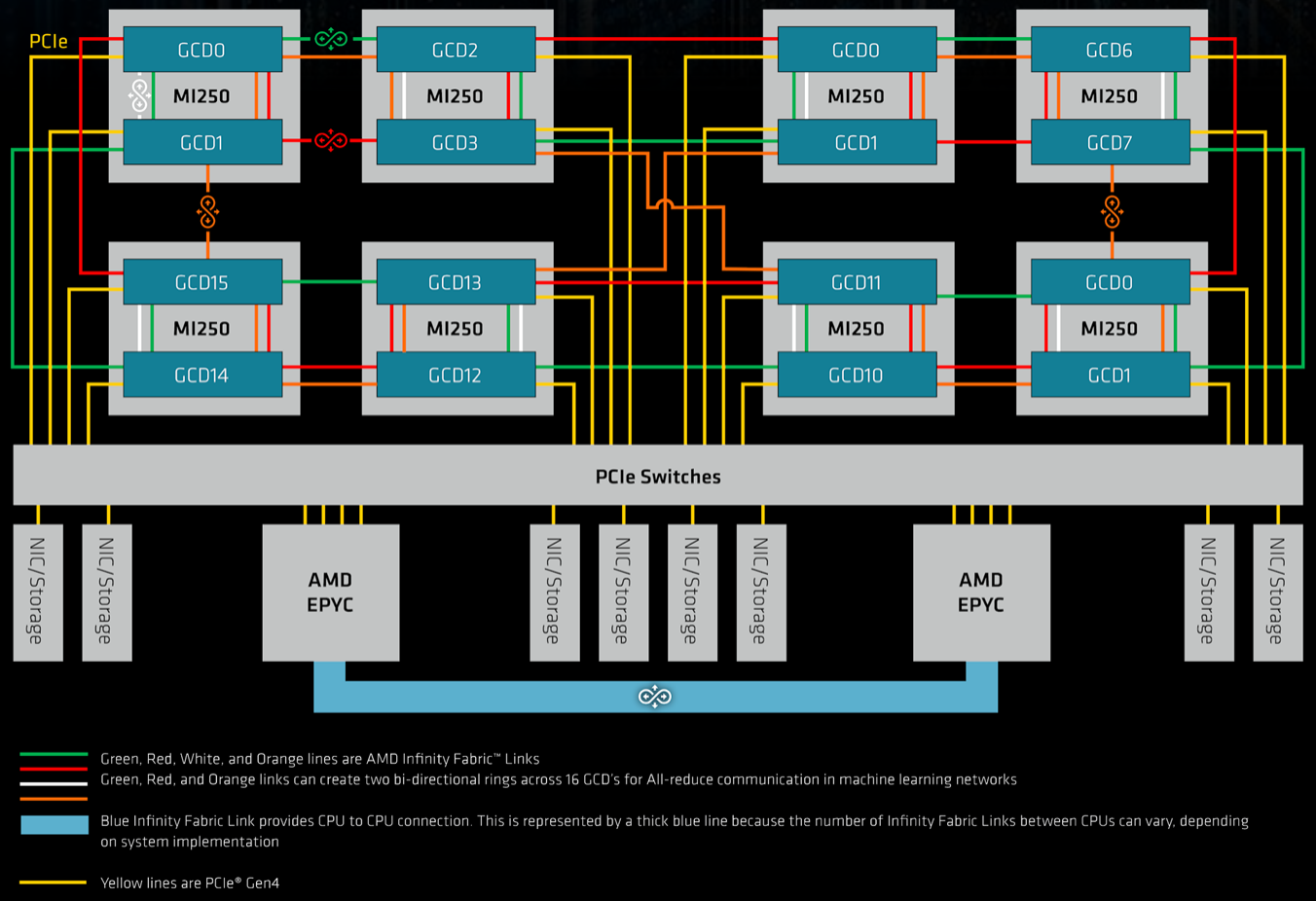
Four AMD Instinct
MI250 Accelerators

5 GPU-to-GPU
Infinity Fabric Links

2 PCIe® Gen 4
CPU-to-GPU Links Per OAM

PCIe® Switches
for NIC RDMA

Flagship Machine Learning Topology with AMD Instinct™ MI250 GPUs



Two 3rd Gen
AMD EPYC™ Processors

Eight AMD Instinct
MI250 Accelerators

5 GPU-to-GPU
Infinity Fabric Links

2 PCIe® Gen 4
CPU-to-GPU Links Per OAM

PCIe® Switches
for NIC RDMA

GIGABYTE™

G262-ZO0 Server

Dual AMD EPYC™ CPU

16 DIMMs at 3200MHz

4x 2.5" Gen4 U.2 NVMe/SATA/SAS Hot-swap Bays

6x low-profile PCIe® Gen4 x16 slots

1x OCP 3.0 Gen4 x16 Mezzanine Slot

Four AMD Instinct™ MI250 GPUs at 560W

2U High

AMD

AS-4124GQ-TNMI Server

Four AMD Instinct™ MI250 GPUs at 530W

10 Hot-swap 2.5" U.2 NVMe/SATA/SAS Hybrid Drive Bays

8 slots PCIe® 4.0 x16 (Low-Profile) via PCIe Switch

Dual AMD EPYC™ 7003 Series CPUs

32 DIMMs, up to 8TB Registered ECC DDR4-3200MHz

4U High with 4x 3000W Titanium Level Redundant Power Supplies

Summary

Domain-specific Architecture
Optimal Efficiency through Domain-specific Optimization

HBM2e Memory System
128 GB per MI250X

3rd Generation AMD Infinity Architecture
Unified Shared Memory

Powerful and Efficient
Number 1 on Top500 and Top 4 on Green500

Endnotes

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

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

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

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

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

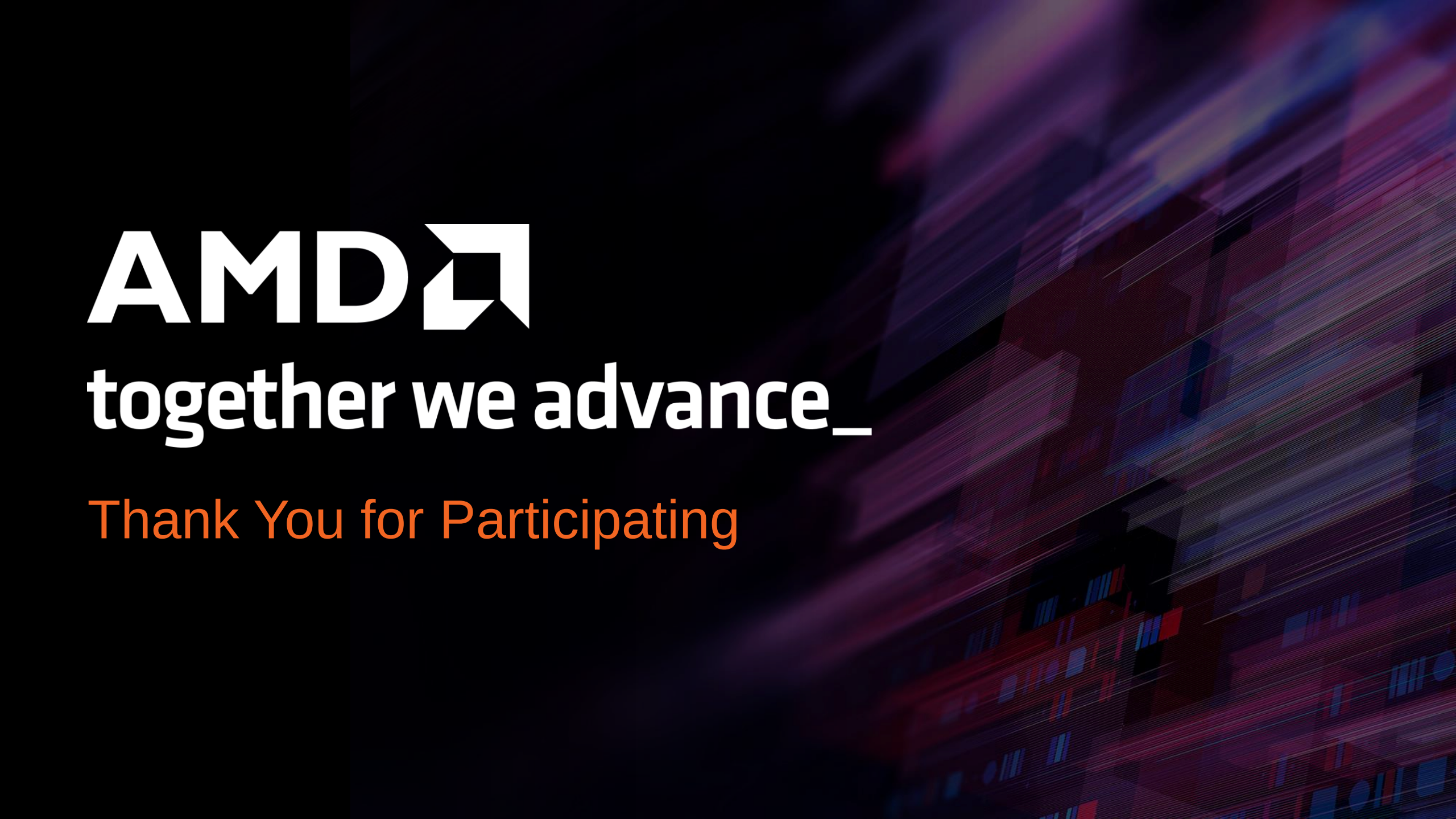
MI200-64: AMD testing as of 8/3/2021 (MI250X) and 1/20/2021 (MI100) using DGEMM 4K Trig Float Init, DPM on. Systems compared: 2P Optimized 3rd Gen EPYC CPUs with 4x AMD Instinct MI250X (560W, 220CUs) running ROCm 4.3.1-59 vs. 1P EPYC 7742 with 1x AMD Instinct MI100 (300W, 120CUs) running ROCm 3.7.0-3289.

GROMACS: http://www.gromacs.org/

HACC: https://cpac.hep.anl.gov/projects/hacc/

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

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


AMD
together we advance_
Thank You for Participating