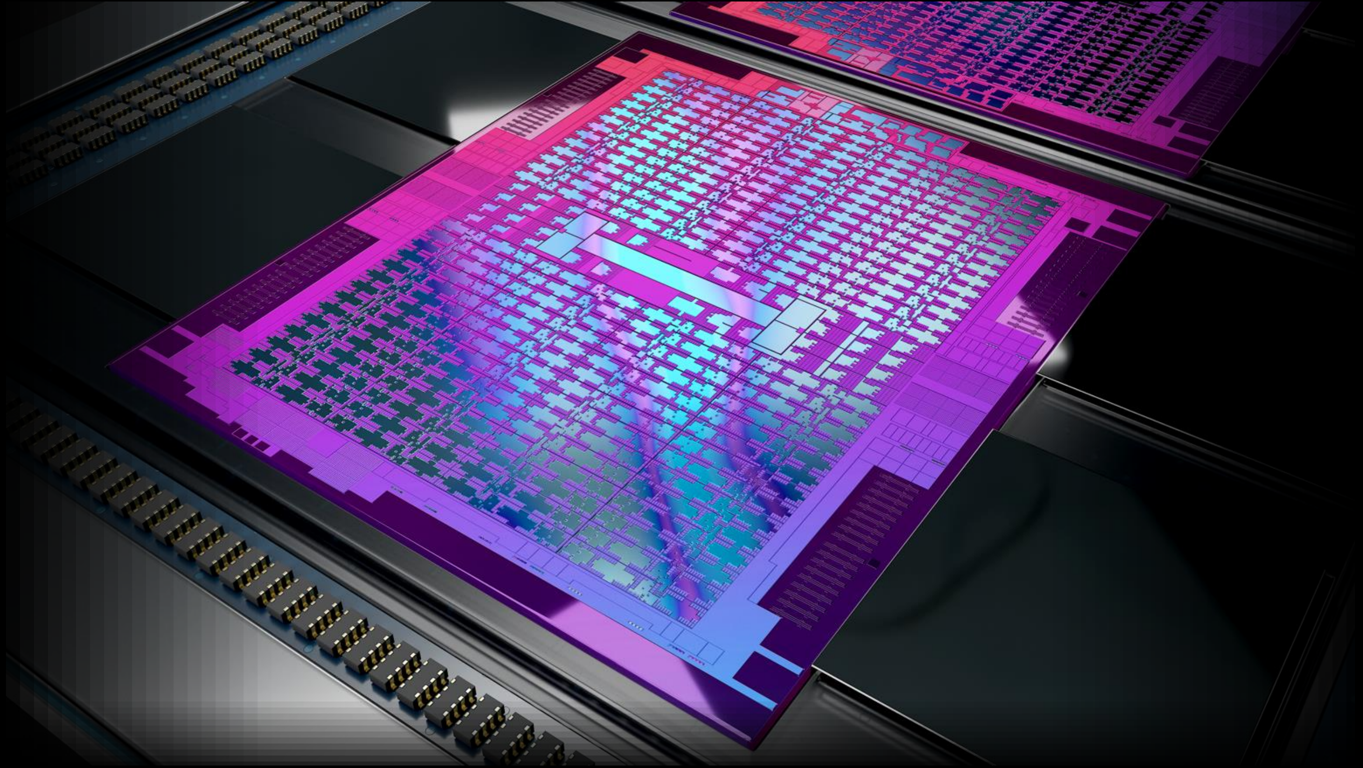
HOW AMD PAVED THE PATH TO EXASCALE

Simón Viñals, Commercial Sales Manager, Spain & Portugal @ AMD
simon.vinals@amd.com

Sept. 14th, 2022
JURES 22, Cáceres, Spain

[redacted]



THE WORLD CAN'T ACCELERATE ENOUGH

RESEARCHERS
must solve the greatest problems in history *faster*

AGENCIES AND ACADEMIC INSTITUTIONS
must *speed* achievement of better lives, economies, communities

THE ENTERPRISE
must leverage HPC and AI innovations *quickly* and profitably

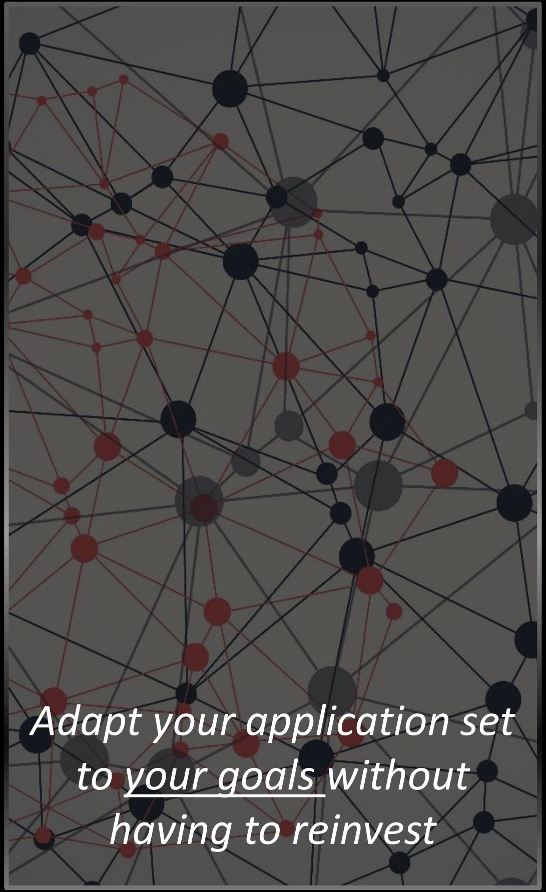
BUT… RESOURCES ARE LIMITED AND ARCHITECTURES ARE RIGID

WHAT DO YOU NEED TO SPEED OUTCOMES IN HPC AND AI?

EFFICIENCY

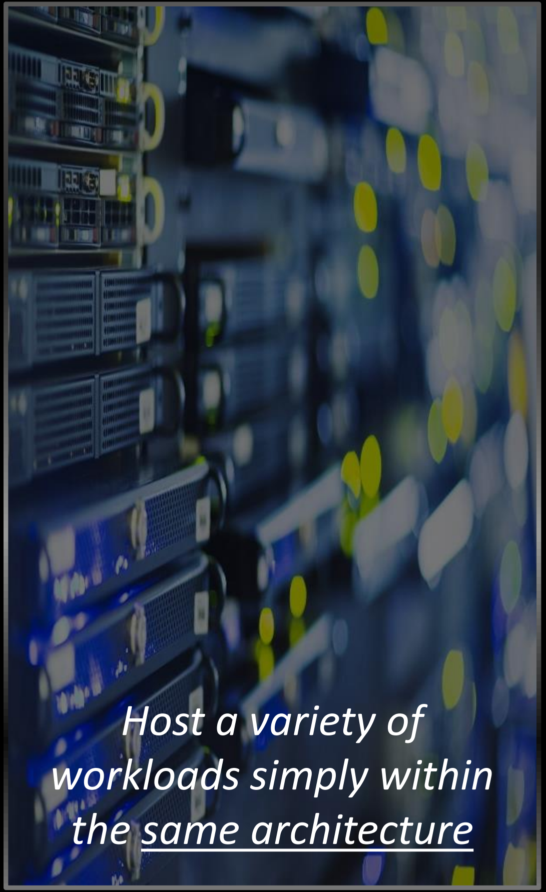
Maximize performance per watt to ease power limitations

PLATFORM FLEXIBILITY



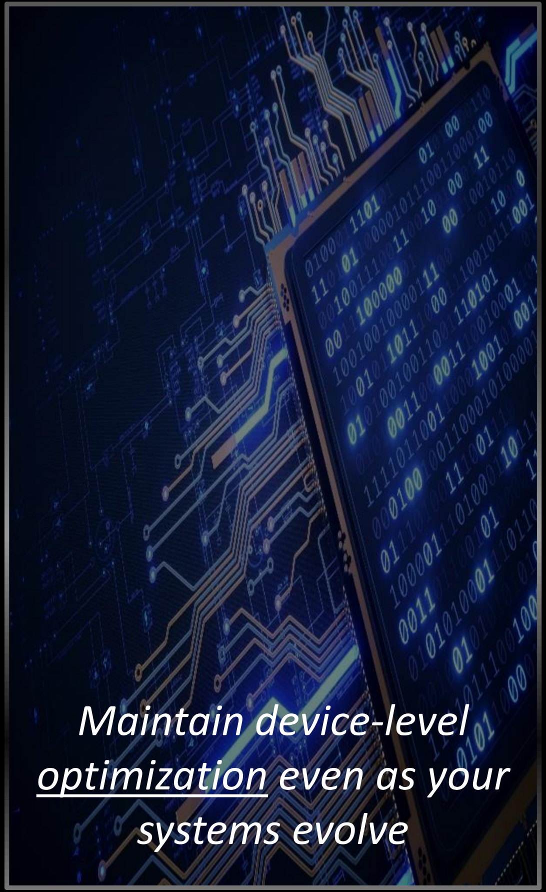
Adapt your application set to your goals without having to reinvest

VERSATILE PERFORMANCE



Host a variety of workloads simply within the same architecture

CODE READINESS



Maintain device-level optimization even as your systems evolve

AMD

EPYC | INSTINCT

LEADING THE EXASCALE ERA

- Powering World's #1 Supercomputer
  First to break Exascale barrier
- Powering World's #1 Green Supercomputer
  8 of top 10 most efficient systems rely on AMD
- Powering World's #1 AI Supercomputer
  More than 3X the previous record holder
- 95% growth in TOP500 systems Year-over-Year
  Powering more than half of all new systems

CRAY
OAK RIDGE National Laboratory
U.S. DEPARTMENT OF ENERGY
Hewlett Packard Enterprise
AMD
FRONTIER

* HPL score achieved 5/30/22 by the Frontier system at Oak Ridge National Laboratory. TOP500 https://www.top500.org/news/ornls-frontier-first-to-break-the-exaflop-ceiling

WORLDS FIRST EXASCALE SUPERCOMPUTER – 1.1 EXAFLOP

TOP 500 The List.

#1 FRONTIER

#3 LUMI

#10 GENCI Adastra CINES

THE GREEN 500

#1 OAK RIDGE National Laboratory Frontier TDS

#2 FRONTIER

#3 LUMI

#4 GENCI Adastra CINES

OAK RIDGE National Laboratory | Lawrence Livermore National Laboratory | CSC | VETENSKAP OCH KONST Kungl Tekniska Högskolan | PAWSEY supercomputing centre | GENCI | Nikhef | CGG | eni | ARL | NORTHERN DATA | NCSA

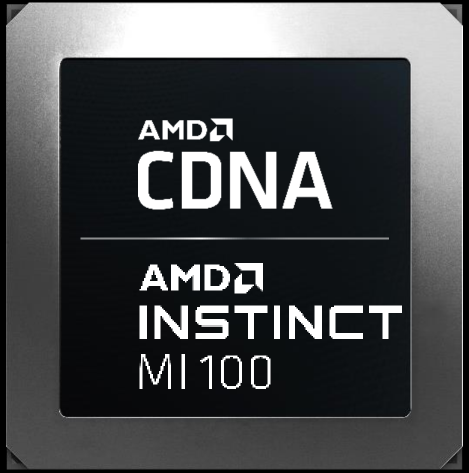
AMD together we advance_

COMPUTE GPU ARCHITECTURE ROADMAP

2020

✓

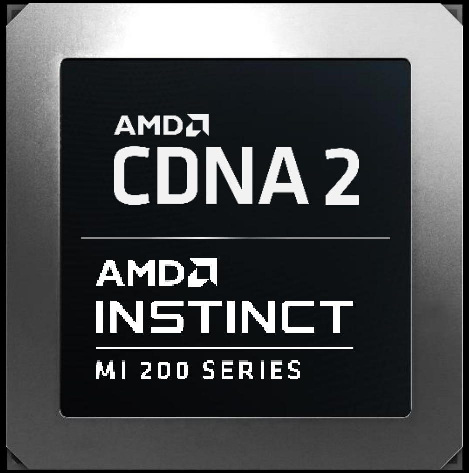
2H'20



7nm

✓

2H21



6nm

2H23

5nm

2023

Roadmaps Subject to Change

AMD INSTINCT™ MI200 SERIES

TACKLING YOUR MOST COMPLEX AND COMPUTE-INTENSIVE PROBLEMS

AMD INSTINCT™

MI200 OAM SERIES

MI250, MI250X

AMD INSTINCT™

MI200 PCIe® SERIES

MI210

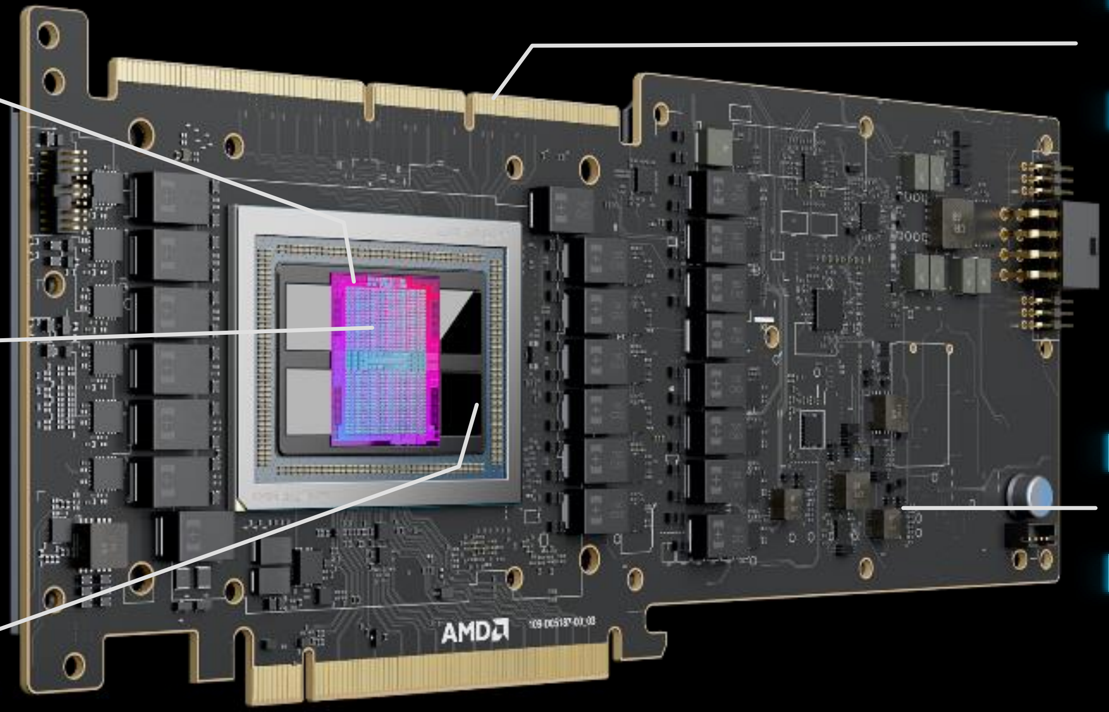
SHATTERING PERFORMANCE BARRIERS IN HPC & AI

DELIVERED PERFORMANCE	A100	MI250	INSTINCT ADVANTAGE
FP64 VECTOR	9.5 TF	30.0 TF	3.1X
FP32 VECTOR	19.5 TF	42.0 TF	2.1X
PACKED FP32 VECTOR	N/A	67.0 TF	N/A
FP64 MATRIX	19.0 TF	54.5 TF	2.8X
FP32 MATRIX	N/A	67.0 TF	N/A
FP16 MATRIX	290 TF	298 TF	1.0X
MEMORY SIZE	80 GB	128 GB	1.6X
PEAK MEMORY BANDWIDTH	2.0 TB/s	3.2 TB/s	1.6X

NOTE: THE A100 TF32 DATA FORMAT IS NOT IEEE FP32 COMPLIANT , SO NOT INCLUDED IN THIS COMPARISON.

SEE ENDNOTES: MI200-01, MI200-07

AMD INSTINCT™ MI210 ACCELERATOR



AMD CDNA™ 2
GRAPHICS COMPUTE DIE

2nd GEN MATRIX CORES

64GB HBM2e @ 1.6 TB/s

2-WAY & 4-WAY INFINITY
FABRIC™ BRIDGES

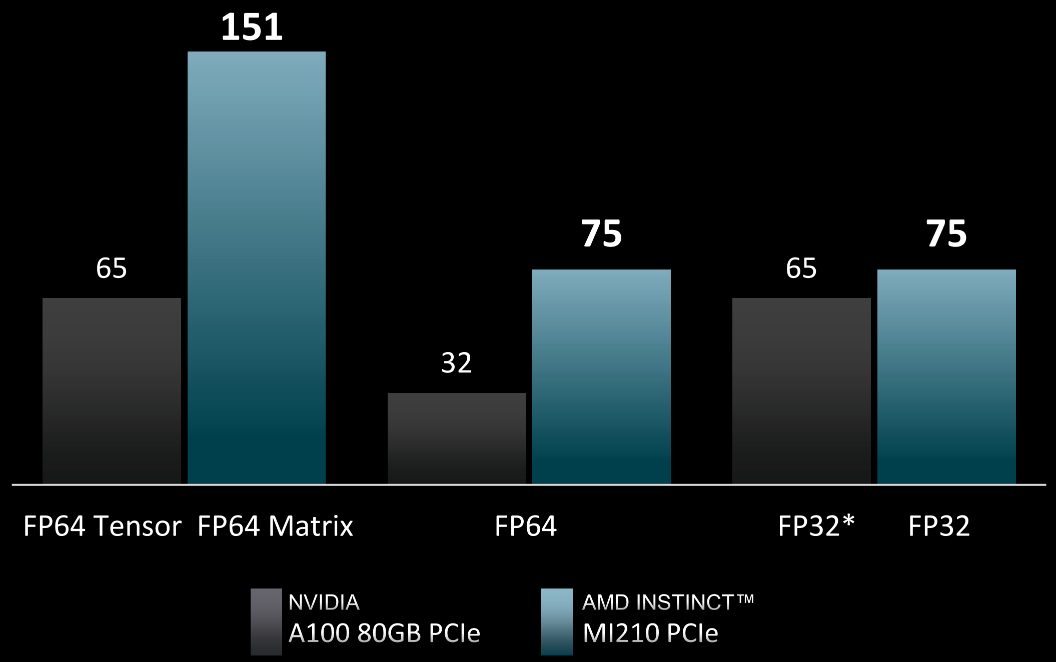
SR-IOV SUPPORT
FOR VMWARE®

AMD INSTINCT™ MI210 ACCELERATOR
POWERFUL & EFFICIENT PCIE® GPU

Leadership HPC Performance
(Peak TFLOPS)

	A100 PCIe®	MI210	INSTINCT™ ADVANTAGE
FP64	**9.7 TF**	**22.6 TF**	**2.3X**
FP32	**19.5 TF**	**22.6 TF**	**1.2X**
FP64 (TENSOR vs. MATRIX)	**19.5 TF**	**45.3 TF**	**2.3X**
FP32 MATRIX	**N/A**	**45.3 TF**	N/A

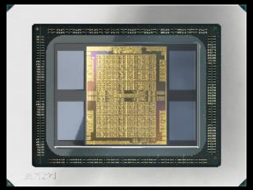
Leadership Efficiency
(GFLOPS/Watt)



*THE A100 TF32 DATA FORMAT IS NOT IEEE FP32 COMPLIANT , SO NOT INCLUDED IN THIS COMPARISON.

SEE ENDNOTES: MI200-41, MI200-44

OUR JOURNEY IN GPU ACCELERATION

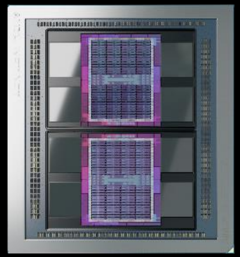


AMD Instinct™ MI100

AMD CDNA™

Ecosystem Growth

First purpose-built GPU architecture for the data center



AMD Instinct™ MI200

AMD CDNA™ 2

Driving HPC and AI to a New Frontier

First purpose-built GPU powering discovery at Exascale

AMD Instinct™ MI300

AMD CDNA™ 3

Data Center APU

Breakthrough architecture designed for leadership efficiency and performance for HPC and AI

2020 → 2023

Roadmaps Subject to Change

AMD INSTINCT™ MI300

THE WORLD'S FIRST DATA CENTER APU

- 4th Gen AMD Infinity Architecture: AMD CDNA™ 3 and EPYC™ CPU "Zen 4" Together

  CPU and GPU cores share a unified on-package pool of memory

- Groundbreaking 3D Packaging

  CPU | GPU | Cache | HBM

- Designed for Leadership Memory Capacity, Bandwidth and Application Latency

- APU Architecture Designed for Power Savings Compared to Discrete Implementation

Available 2023

APU: Accelerated Processing Unit

>8X

Expected AI Training Performance vs. MI250X

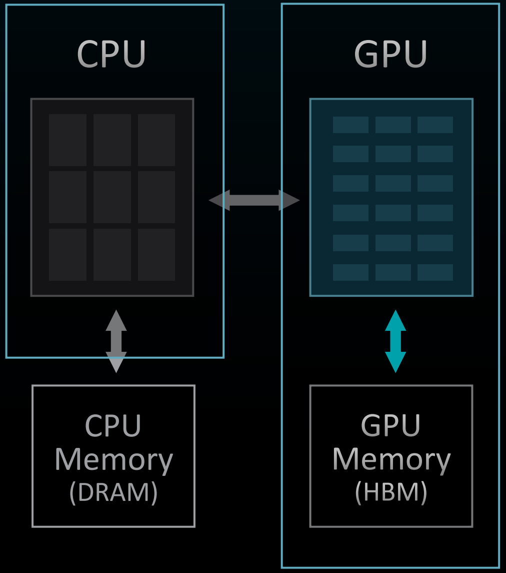
See Endnote MI300-03. Preliminary data and projections, subject to change.

AMD
together we advance_

UNIFIED MEMORY APU ARCHITECTURE BENEFITS

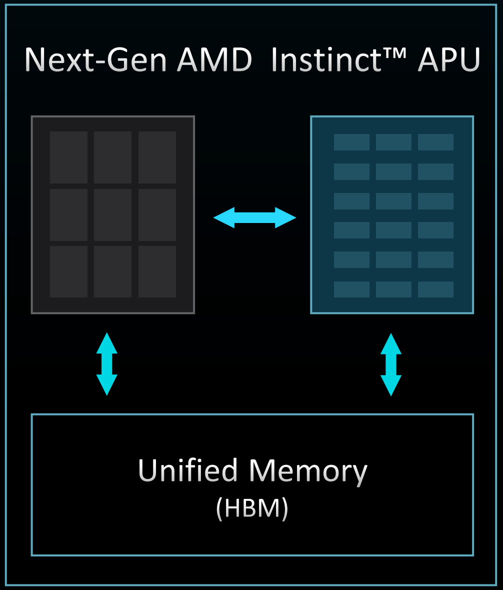
AMD CDNA™ 2 Coherent Memory Architecture

- **Simplifies Programming**
- **Low Overhead 3rd Gen Infinity Interconnect**
- **Industry Standard Modular Design**



AMD CDNA™ 3 Unified Memory APU Architecture

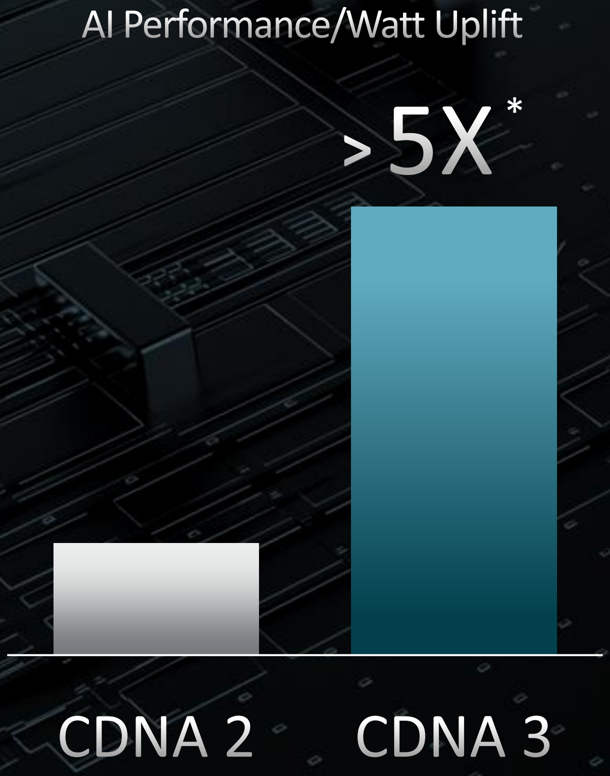
- **Eliminates Redundant Memory Copies**
- **High-Efficiency 4th Gen Infinity interconnect**
- **Low TCO with Unified Memory APU Package**



*AMD Projections

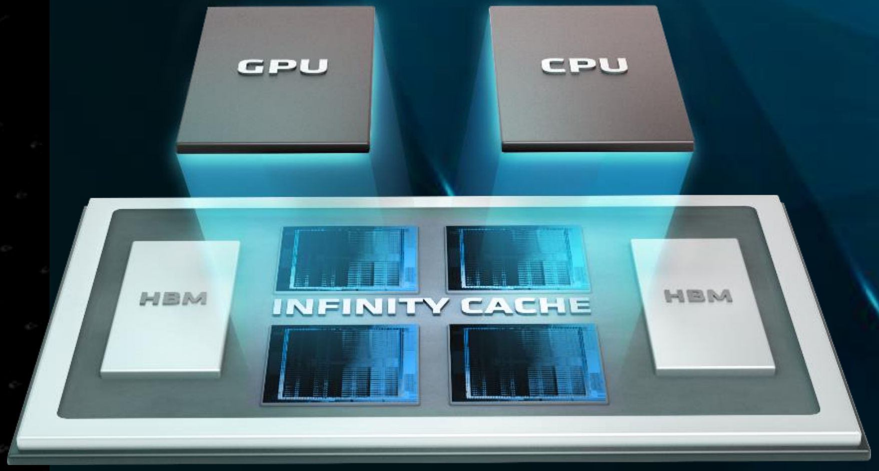
AMD
CDNA 3

THE JOURNEY CONTINUES



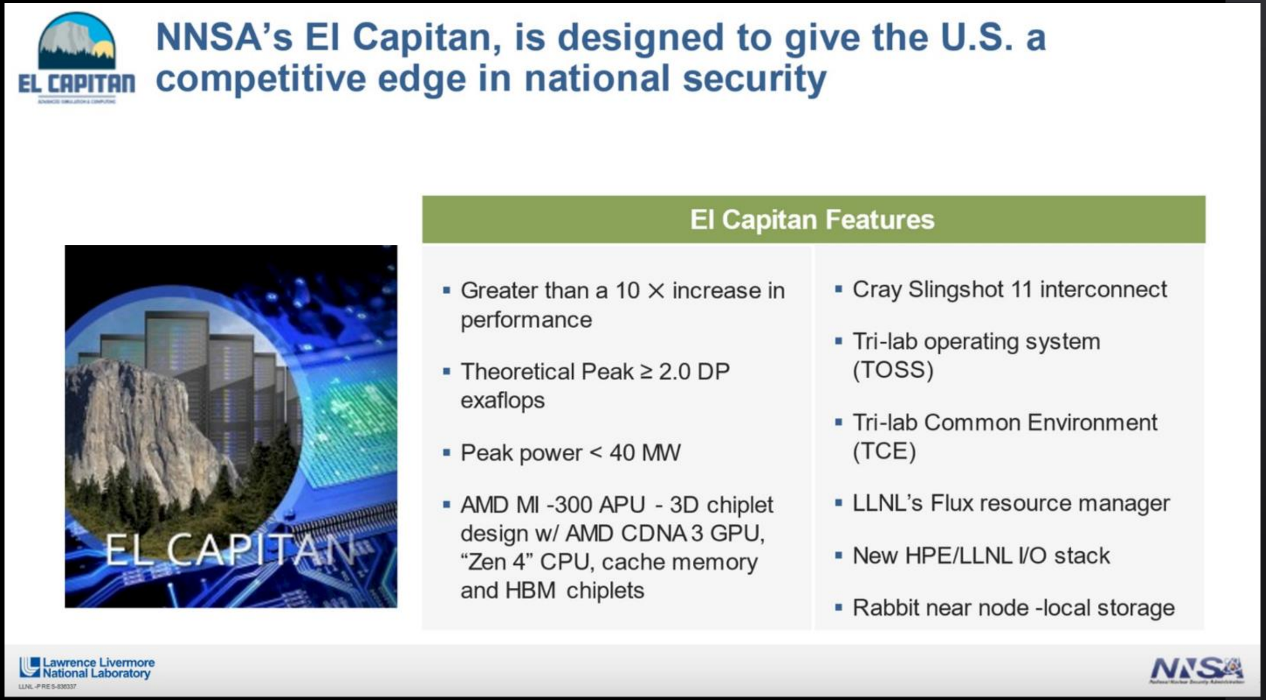
EXPECTED PERFORMANCE-PER-WATT UPLIFT THROUGH:

- 5nm Process and 3D Chiplet Packaging
- Next-Gen AMD Infinity Cache™
- 4th Gen Infinity Architecture
- Unified Memory APU Architecture
- New Math Formats



*See endnote MI300-04. Preliminary data and projections, subject to change

LEADING THE EXASCALE ERA AGAIN



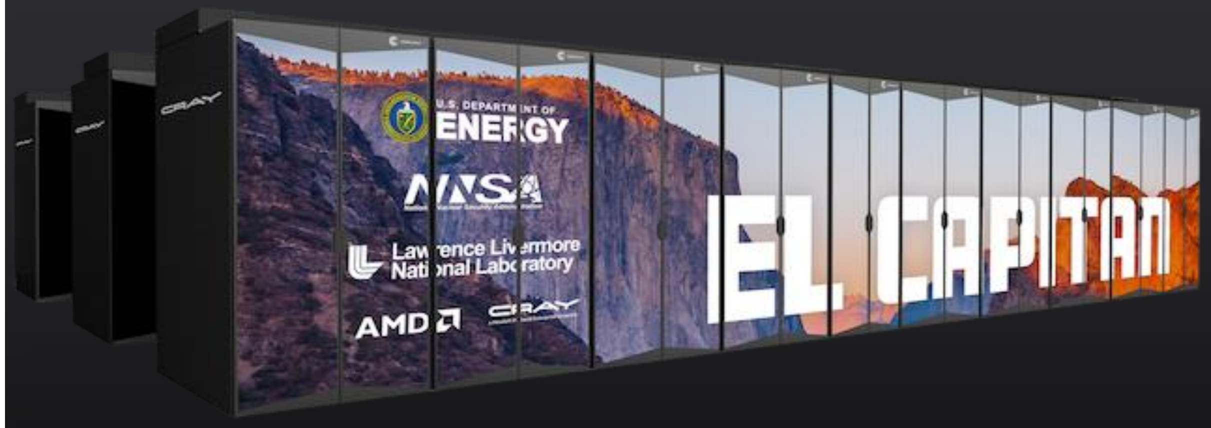
NNSA's El Capitan, is designed to give the U.S. a competitive edge in national security

El Capitan Features

- Greater than a 10 × increase in performance
- Theoretical Peak ≥ 2.0 DP exaflops
- Peak power < 40 MW
- AMD MI -300 APU - 3D chiplet design w/ AMD CDNA 3 GPU, "Zen 4" CPU, cache memory and HBM chiplets
- Cray Slingshot 11 interconnect
- Tri-lab operating system (TOSS)
- Tri-lab Common Environment (TCE)
- LLNL's Flux resource manager
- New HPE/LLNL I/O stack
- Rabbit near node -local storage

Lawrence Livermore National Laboratory

LLNL-PRES-836337


U.S. DEPARTMENT OF ENERGY
NNSA
Lawrence Livermore National Laboratory
AMD
CRAY
EL CAPITAN

Expected Late 2023

TOP500, Green500, and HPL-AI lists, as of May 30,2022

AMD ROCm 5.0

DEMOCRATIZING EXASCALE FOR ALL

EXPANDING SUPPORT & ACCESS

OPTIMIZING PERFORMANCE

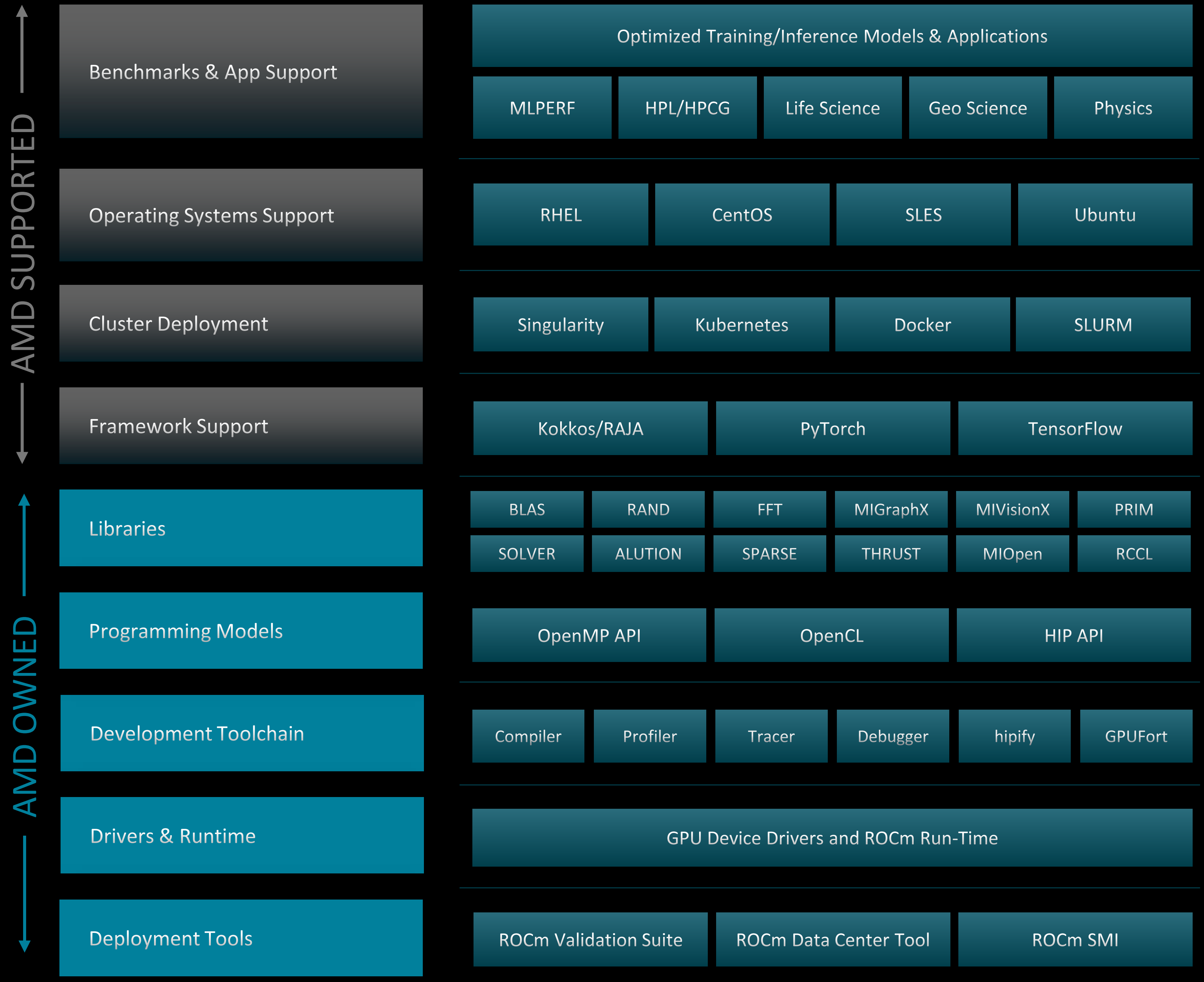
ENABLING DEVELOPER SUCCESS

AMD together we advance_

OPEN SOFTWARE PLATFORM

AMD ROCm

- Unlocked GPU power to accelerate computational tasks
- Optimized for HPC and Deep Learning workloads at scale
- Open source enabling innovation, differentiation, and collaboration



AMD SUPPORTED	Benchmarks & App Support	Optimized Training/Inference Models & Applications MLPERF, HPL/HPCG, Life Science, Geo Science, Physics
	Operating Systems Support	RHEL, CentOS, SLES, Ubuntu
	Cluster Deployment	Singularity, Kubernetes, Docker, SLURM
	Framework Support	Kokkos/RAJA, PyTorch, TensorFlow
AMD OWNED	Libraries	BLAS, RAND, FFT, MIGraphX, MIVisionX, PRIM SOLVER, ALUTION, SPARSE, THRUST, MIOpen, RCCL
	Programming Models	OpenMP API, OpenCL, HIP API
	Development Toolchain	Compiler, Profiler, Tracer, Debugger, hipify, GPUFort
	Drivers & Runtime	GPU Device Drivers and ROCm Run-Time
	Deployment Tools	ROCm Validation Suite, ROCm Data Center Tool, ROCm SMI

AMD together we advance_

ML FRAMEWORKS & LIBRARIES

UPSTREAMED SOURCE & BINARY SUPPORT
ALLOW SCIENTISTS TO EASILY USE EXISTING CODE

	Source	Container	PIP Wheel
TensorFlow	TensorFlow GitHub	Infinity Hub	pypi.org
PyTorch	PyTorch GitHub	Infinity Hub	pytorch.org
ONNX RUNTIME	ONNX-RT GitHub	Docker Instructions	onnxruntime.ai
JAX	GitHub public fork	Docker Hub	Est 2022
DeepSpeed	DeepSpeed GitHub	Docker Hub	deepspeed.ai
CuPy	cupy.dev	Docker Hub	cupy.dev

BROAD APPLICATION SUPPORT

Application catalog growing rapidly

Instinct Application Catalog

Application Name	Category	MI100/MI200
CP2K	Quantum Chemistry	Now
AMBER*	Life Science	Now
GROMACS	Life Science	Now
NAMD	Life Science	Now
OpenMM	Life Science	Now
LAMMPS	Life Science	Now
Relion	Life Science	Now
SPECFEM3D - Cartesian	CFD	Now
SPECFEM3D - Globe	CFD	Now
GRID (QCD)	Physics	Now
MILC**	Physics	Now
Chroma**	Physics	Now
GRID (CPS)	Physics	Now
LSMS	Physics	Now
Mini-HACC	Cosmology	Now
TensorFlow	Machine Learning	Now
PyTorch	Machine Learning	Now
PyFR	Machine Learning	Now
HPL	Benchmark	Now
HPCG	Benchmark	Now
AMG (Setup/Solve)	Benchmark	Now

Application Name	Category	MI100 / MI200
Nbody (32/64)	Benchmark	Now
Quiksilver	Benchmark	Now
Ansys Mechanical*	ISV	2H' 2022
OpenFOAM	CFD	2H' 2022
MPAS	Weather	2H' 2022
ICON	Weather	2H' 2022
VASP	Life Science	2H' 2022
Hoomd-Blue	Life Science	2H' 2022
NWCHEM	Quantum Chemistry	2H' 2022
AceCast (Based on WRF)	ISV	2H' 2022
HPL-AI	Benchmark	2H' 2022
PETSc	Library	2H' 2022
SHOC	Benchmark	2H' 2022
BabelStream	Benchmark	2H' 2022
Cholla	CFD	2H' 2022
Relion v4	Cryo-EM	2H' 2022
PIConGPU	Physics	2H' 2022
DL-Poly	Molecular Dynamics	1H' 2023
VTKm	Visualization	1H' 2023
Quantum Espresso	Physics	1H' 2023
GPAW	Physics	1H' 2023
GADGET(v4)	Cosmology	1H' 2023

*Commercial SW with links to ported/optimized code or application
**Two containers, one for MI100 and one for MI200

together we advance_

CODE CONVERSION TOOLS

EXTEND YOUR APPLICATION PLATFORM SUPPORT BY CONVERTING CUDA® CODE TO ROCM

HIPIFY-PERL

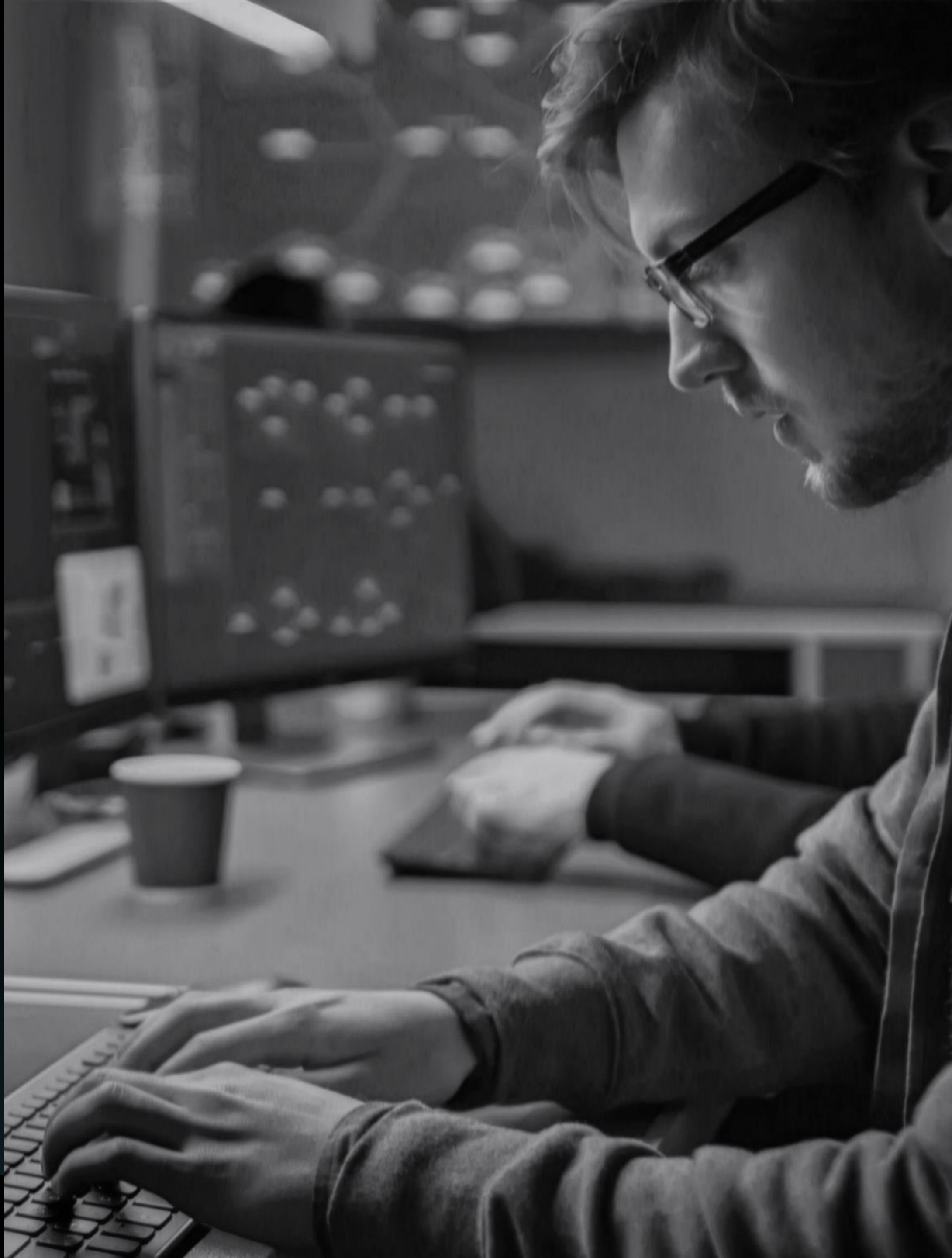
- Easiest to use; point at a directory and it will hipify CUDA code
- Very simple string replacement technique; may require manual post-processing
- Recommended for quick scans of projects

HIPIFY-CLANG

- More robust translation of the code
- Generates warnings and assistance for additional analysis
- High quality translation, particularly for cases where the user is familiar with the make system

gpuFORT

- Conversion tool to translate directive-based code to direct kernel programing source code – early release available on github
- Fortran + OpenACC and CUDA Fortran convert to:
  - Fortran + [GCC/AOMP OpenACC/MP runtime calls] + HIP C++
  - Fortran + OpenMP 4.5+



ARE YOU READY TO

ACCELERATE DISCOVERY?

Learn more about AMD Instinct™ MI200 Accelerators and ROCm™ 5 Powering Discoveries at Exascale

AMD.COM/INSTINCT

AMD
together we advance_

DISCLAIMER AND ATTRIBUTIONS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18

ENDNOTES

MI200-01 - World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-01

MI200-40 - "Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) 500 Watt accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64) floating-point performance. The Nvidia A100 SXM (80 GB) accelerator (400W) with boost engine clock of 1410 MHz results in 19.5 TFLOPS peak theoretical double precision (FP64 Tensor Core), 9.7 TFLOPS peak theoretical double precision (FP64) floating-point performance. MI250X 95.7 TFLOPS FP64 Matrix divided by 500 Watts = 0.1914 TFLOPS (191 GFLOPS) per watt.MI250X 47.9 TFLOPS FP64 divided by 500 Watts = 0.0958 TFLOPS (96 GFLOPS) per watt.A100 19.5 TFLOPS FP64 Tensor Core divided by 400W = .0488 TFLOPS (49 GFLOPS) per watt.A100 9.7 TFLOPS FP64 divided by 400W = .0243 TFLOPS (24 GFLOPS) per watt..1914/.0488= 3.9x the/2.9x better peak perf/watt.0958/.0243=3.9x the /2.9x better peak perf/watt"

MI200-41 - Calculations conducted by AMD Performance Labs as of Jan 14, 2022, for the AMD Instinct™ MI210 (64GB HBM2e PCIe® card) accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), and 181.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), and 184.6 TFLOPS peak theoretical half precision (FP16), floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64) and 39 TFLOPS peak Bfloat16 format precision (BF16), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-41

MI200-42 - 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

MI200-44 - Calculations conducted by AMD Performance Labs as of Feb 15, 2022, for the AMD Instinct™ MI210 (64GB HBM2e PCIe® card) 300 Watt accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), 22.6 TFLOPS peak theoretical single precision (FP32), and 181.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), floating-point performance. AMD TFLOPS calculations conducted with the following equation for AMD Instinct MI210 and MI100 GPUs: FLOPS calculations are performed by taking the engine clock from the highest DPM state and multiplying it by xx CUs per GPU. Then, multiplying that number by xx stream processors, which exist in each CU. Then, for MI210 that number is multiplied by 256 FLOPS per clock/CU for FP64 Matrix to determine TFLOPS, 128 FLOPS per clock/CU for FP64 to determine TFLOPS, 128 FLOPS per clock/CU for FP32 to determine TFLOPS, 1024 FLOPS per clock/CU for BF16 to determine TFLOPS. Divide results by 100,000 to get TFLOPS. Then, for MI100 that number is multiplied by 64 FLOPS per clock/CU for FP64 to determine TFLOPS. Divide results by 100,000 to get TFLOPS. Published results on the NVidia Ampere A100 (80GB) 300 Watt PCIe® GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 39 TFLOPS peak Bfloat16 format precision (BF16), theoretical floating-point performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1.

MI210 45.3 TFLOPS FP64 Matrix divided by 300 Watts = 0.151 TFLOPS (151 GFLOPS) per watt.; MI210 22.6 TFLOPS FP64 divided by 300 Watts = 0.0753 TFLOPS (75 GFLOPS) per watt.; MI210 22.6 TFLOPS FP32 divided by 300 Watts = 0.0753 TFLOPS (75 GFLOPS) per watt.; MI210 181.0 TFLOPS BF16 divided by 300 Watts = 0.6033 TFLOPS (603 GFLOPS) per watt.; A100 19.5 TFLOPS FP64 Tensor Core divided by 300W = .065 TFLOPS (65 GFLOPS) per watt.; A100 9.7 TFLOPS FP64 divided by 300W = .0323 TFLOPS (32 GFLOPS) per watt.; A100 19.5 TFLOPS FP32 divided by 300W = .065 TFLOPS (65 GFLOPS) per watt.; A100 39 TFLOPS BF16 divided by 300W = .13 TFLOPS (130 GFLOPS) per watt.

.151/.065 = 2.3x the/1.3x better peak perf/watt (FP64 Matrix)

.0753/.0323 = 2.3x the /1.3x better peak perf/watt (FP64)

.0753/.065 = 1.15x the/0.15x better peak perf/watt (FP32)

0.6033/.13 = 4.6x the/3.6 better peak perf/watt (BF16)

ENDNOTES (CONT.)

MI200-46 – Testing Conducted by AMD performance lab on a 2P socket 3rd Gen AMD EPYC™ '7763 CPU Supermicro 4124, with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W) No AMD Infinity Fabric™ technology enabled. OS: Ubuntu 18.04.6 LTS, ROCm 5.0 Benchmark: Relion v3.1.2 Converted with HIP with AMD optimizations to Relion that are not yet available upstream. Vs. Nvidia Published Measurements: https://developer.nvidia.com/hpc-application-performance accessed 3/13/2022 Nvidia Container details found at: https://catalog.ngc.nvidia.com/orgs/hpc/containers/relion information on Relion found at: https://www2.mrc-lmb.cam.ac.uk/relion/index.php/Download_%26_install All results measured on systems with 8 GPUs, with 4GPU XGMI bridges where applicable. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-46

MI200-47 – Testing Conducted by AMD performance lab on a 2P socket 3rd Gen AMD EPYC™ '7763 CPU Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W) No AMD Infinity Fabric™ technology enabled. OS: Ubuntu 18.04.6 LTS, ROCm 5.0 Benchmark: LAMMPS ReaxFF/C, patch_2Jul2021 plus AMD optimizations to LAMMPS and Kokkos that are not yet available upstream Metric: ATOM timesteps per second Vs. Nvidia published measurement with 8x NVIDIA A100 GPU (PCIe 80GB 300W) using benchmark LAMMPS Reaxff Atom Timesteps/s (atom_timesteps/s) (stable_29Sep2021) Benchmark: LAMMPS classical molecular dynamics package ReaxFF/C, patch_10Feb2021 resulted in a published score of 11,400,000 (1.14E+07) ATOM-Time Steps/s. https://developer.nvidia.com/hpc-application-performance accessed 3.13.2022 Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:lammps Information on LAMMPS: https://www.lammps.org/index.html All results measured on systems with 8 GPUs, with 4GPU XGMI bridges where applicable Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-47

MI200-48 – Testing Conducted by AMD performance lab on a 2P socket 3rd Gen AMD EPYC™ 7763 CPUs Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W), No AMD Infinity Fabric™ technology enabled. Results shown are calculated from the Geomean of 5 runs. OS: Ubuntu 18.04.6 LTS, ROCm 5.0 Benchmark Quicksilver, Hipified version of a proxy app for the Monte Carlo Transport Code, Mercury. LLNL-CODE-684037 plus AMD optimizations that are on AMD Github branch on github. Vs. Testing Conducted by AMD performance on 2P socket AMD EPYC™ 7763 Supermicro 4124 with 8x NVIDIA A100 GPUs (PCIe 40GB 250W) OS: Ubuntu 18.04.6 LTS Benchmark: Quicksilver - LLNL-CODE-684037 run with CUDA code version 11.2.152 All results measured on systems with 8 GPUs, with 4GPU XGMI bridges where applicable. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-48

MI200-49 – Testing Conducted by AMD performance lab on a 2P socket AMD EPYC™ '7763 CPU Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB,300W), AMD Infinity Fabric™ technology enabled. Results calculated from medians of five runs. OS: Ubuntu 18.04.6 LTS ROCm 5.0 with OpenMPI v4 and UCX v1 Benchmark: HPL v2.3 Benchmark Results: HPL, plus AMD optimizations to HPL that are not yet upstream. Vs. Testing Conducted by AMD performance on 2P socket AMD EPYC™ 7763 Supermicro 4124 with 8x NVIDIA A100 GPU (PCIe 40GB 250W) OS: Ubuntu 18.04.6 LTS CUDA 11.5 Benchmark: HPL Nvidia container image 21.4-HPL All results measured on systems configured with 8 GPUs, using 4 GPU AMD Infinity Fabric(TM) link bridges. Information on HPL: https://www.netlib.org/benchmark/hpl/ Nvidia HPL Container Detail: https://ngc.nvidia.com/catalog/containers/nvidia:hpc-benchmarks Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-49

MI200-50 - Testing Conducted by AMD performance lab on a 2P socket AMD EPYC™ '7763 CPU Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W), AMD Infinity Fabric™ technology not enabled. Results calculated on median of 5 runs. OS: Ubuntu 18.04.6 LTS, ROCm 5.0 with OpenMPI v4.0.5 & UCX 1.8.1. Benchmark: AMG (Solve) FOM, AMG branch (cuda, HIP): Parray-analysis-cuda (converted with HIP) plus AMD optimizations to AMG (Solve) that are not yet available upstream. Benchmark: AMG Solve FOM_Setup / Sec Vs. Testing Conducted by AMD performance on 2P socket AMD EPYC™ 7763 Supermicro 4124 with 8x NVIDIA A100 GPUs (PCIe 40GB Board 250W) OS: Ubuntu 18.04.6 LTS, CUDA 11.6 Benchmark: AMG (Solve) FOM, AMG branch (cuda): Parray-analysis-cuda All results measured on systems with 8 GPUs, with dual 4GPU hive AMD Infinity Fabric™ technology enabled. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-50

Mi200-51 - Testing Conducted by AMD performance lab, on a 2P socket 3rd Gen AMD EPYC™ 7763 CPU Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W), No AMD Infinity Fabric™ technology enabled. OS: Ubuntu 18.04.6 LTS with ROCm 5.0. Benchmark: AMG (Set up) FOM2, AMG branch (cuda, HIP): Parray-analysis-cuda (converted with HIP) plus AMD optimizations to AMG (Set up) that are not yet available upstream. Vs. Testing Conducted by AMD performance lab on 2P socket AMD EPYC™ 7763 Supermicro 4124 with 8x NVIDIA A100 GPUs (PCIe 40GB 250W) OS: Ubuntu 18.04.6 LTS, CUDA code version 11.6 Benchmark: AMG (Set up) FOM, AMG branch (cuda): Parray-analysis-cuda All results measured on systems with 8 GPUs. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-51

MI200-52 - Testing Conducted by AMD performance lab on a 2P socket AMD EPYC™ '7763 CPU Supermicro 4124 with 8x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W), No AMD Infinity Fabric™ technology enabled. results shown calculated on medin of 5 runs. OS: Ubuntu 18.04.6 LTS ROCm 5.0 Benchmark: Nvidia Nbody 32 CUDA sample version 11.2.152 converted to HIP Benchmark Results: Nvidia Nbody 32 sample code version 11.2.152, plus AMD optimizations to Nbody 32 that are not yet available upstream. For Nbody64: Benchmark: Nbody 64 CUDA Sample version 11.2.152 converted to HIP Benchmark Results: Nvidia Nbody 64 samples code version 11.2.152, plus AMD optimizations to Nbody 64 that are not yet available upstream. Vs.Testing Conducted by AMD performance on 2P socket AMD EPYC™ '7763 Supermicro 4124 with 8x NVIDIA A100 GPUs (PCIe 40GB 250W) OS: Ubuntu 18.04.6 LTS, CUDA 11.6 Benchmark: Nvidia Nbody 32 sample code version 11.2.152 Benchmark: Nvidia Nbody 64 sample code version 11.2.152 Values of benchmark (FOM Segments / Sec– bigger is better) All results measured on systems with 8 GPUs. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-52

MI200-53 - Testing Conducted by AMD performance lab on a 2P socket 3rd Gen AMD EPYC™ 7763 CPU Supermicro 4124 with 1x AMD Instinct™ MI210 GPU (PCIe® 64GB 300W), No AMD Infinity Fabric™ technology enabled. Results calculated on median scores of 5 runs. OS: Ubuntu 18.04.6 LTS with ROCm 5.0 Benchmark: OpenMM_gbsa v7.7.0, commit 87a02d8 Disable packed math for >=MI200 (converted to HIP) and run at double precision (4 simulations*10,000 steps). Benchmark Results: OpenMM_gbsa plus AMD optimizations to OpenMM_gbsa that are not yet upstream Vs. Testing Conducted by AMD performance on 2P socket 3rd Gen AMD EPYC™ 7763 Supermicro 4124 1x NVIDIA A100 GPU (PCIe 40GB 250W Server BIOS: 2.2 System Bios: 2.2, GPU Bios: 92.00.25.00.08 OS: Ubuntu 18.04.6 LTS, CUDA 11.6 Benchmark: OpenMM_gbsa v7.7.0, commit Public repository: Tag 7.7.0 - commit 130124a3f9277b054ec40927360a6ad20c8f5fa6, git clone -b 7.7 All results measured on systems with 8 GPUs. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-53

MI300-004 – Measurements by AMD Performance Labs June 4, 2022. MI250X (560W) FP16 (306.4 estimated delivered TFLOPS based on 80% of peak theoretical floating-point performance). MI300 FP8 performance based on preliminary estimates and expectations. MI300 TDP power based on preliminary projections. Actual results based on production silicon may vary.