AMD Instinct™ Accelerators and the ROCm™ Platform

Michael Klemm
Principal Member of Technical Staff
HPC Center of Excellence

Derek Bouius
Sr. Product Manager
GPU Compute Software

CAUTIONARY STATEMENT

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as AMD product roadmaps; the features, functionality, performance, availability, timing and expected benefits of AMD products; expected availability, timing, and benefits of supported ROCm™ applications and the AMD Infinity Hub with AMD products; and the momentum of AMD Instinct™ accelerators, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD’s Securities and Exchange Commission filings, including but not limited to AMD’s most recent reports on Forms 10-K and 10-Q.

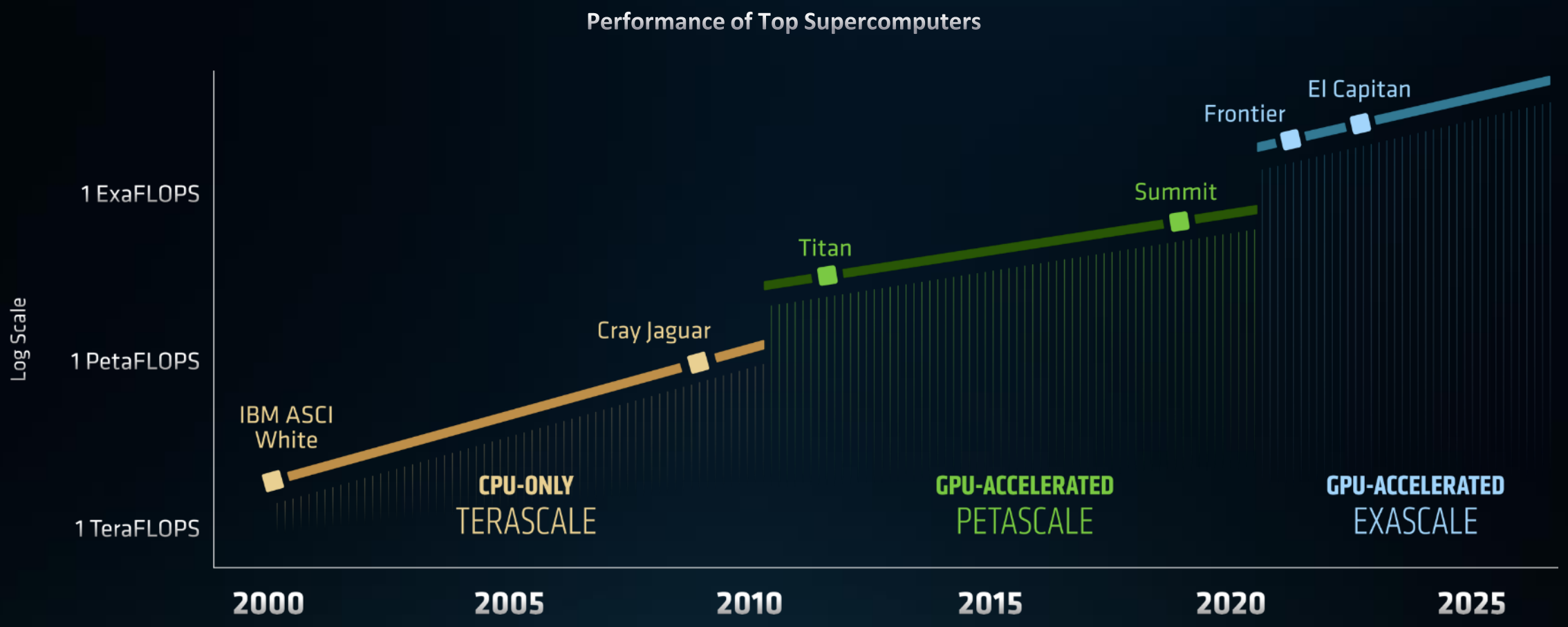
AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

Agenda

- AMD Instinct™ Architecture
- AMD ROCm™ Software Stack
- Ecosystem
- Q&A

AMD Instinct™ Architecture

The Dawn of GPU-Accelerated Exascale:
Major Leaps in Performance Driving Three Phases of Supercomputing



SOURCE: HTTPS://OPENAI.COM/BLOG/AI-AND-COMPUTE/ (MACHINE INTELLIGENCE) AND HTTPS://WWW.TOP500.ORG/ (HIGH PERFORMANCE COMPUTING)

APPLICATION OPTIMIZED ARCHITECTURES

HIGHEST EFFICIENCY THROUGH DOMAIN SPECIFIC OPTIMIZATION

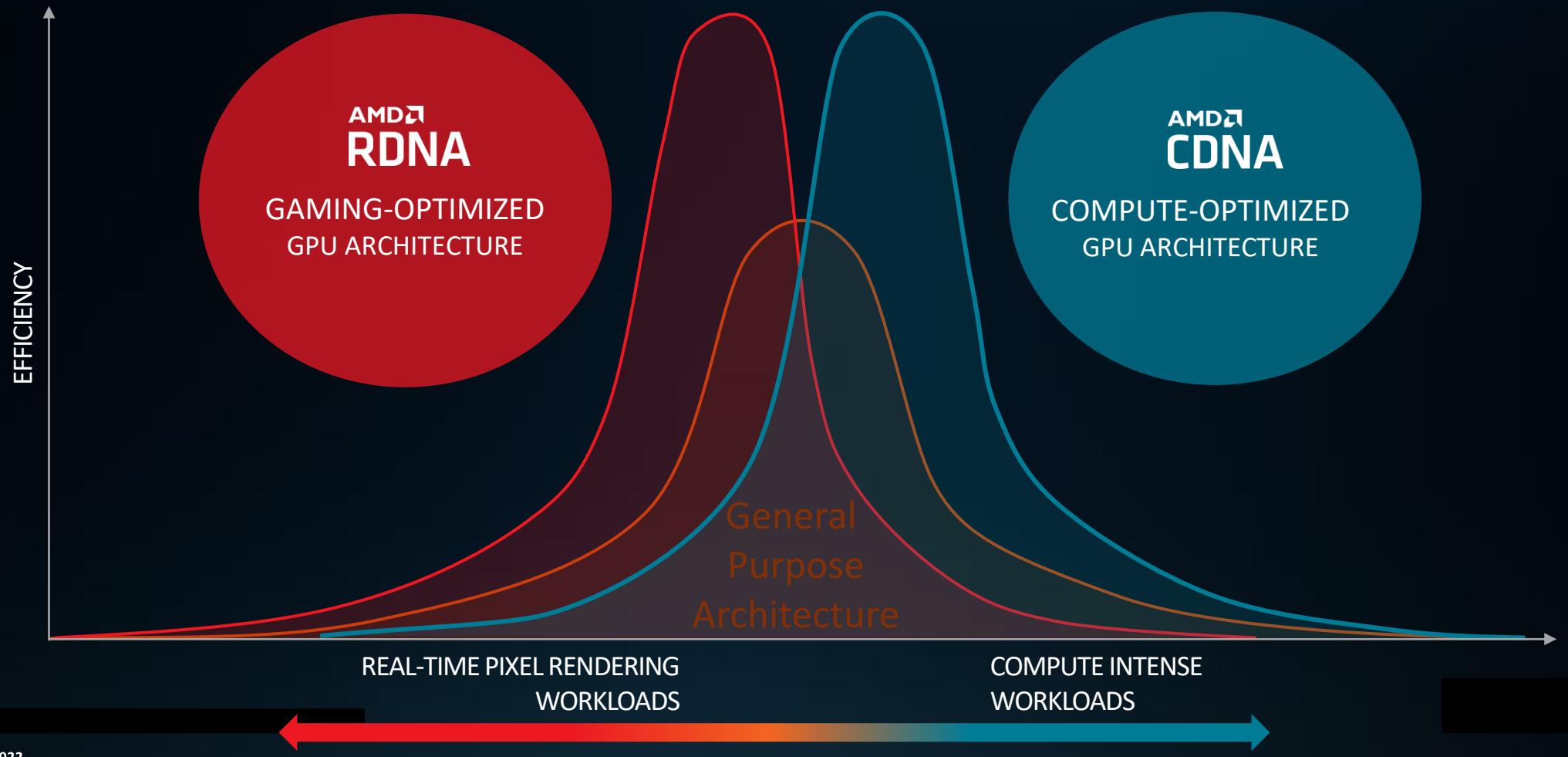


CHART FOR ILLUSTRATIVE PURPOSES

AMD EPYC INSTINCT

LEADING THE NEXT-GEN SUPERCOMPUTING & EXASCALE ERA

VETENSKAP OCH KONST
KUNGL TEKNISKA HÖGSKOLAN

LUMI

CGG GeoConsulting

EuroHPC Joint Undertaking

Nikhef

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

PAWSEY supercomputing centre

Source: https://www.top500.org/lists/top500/2021/06/

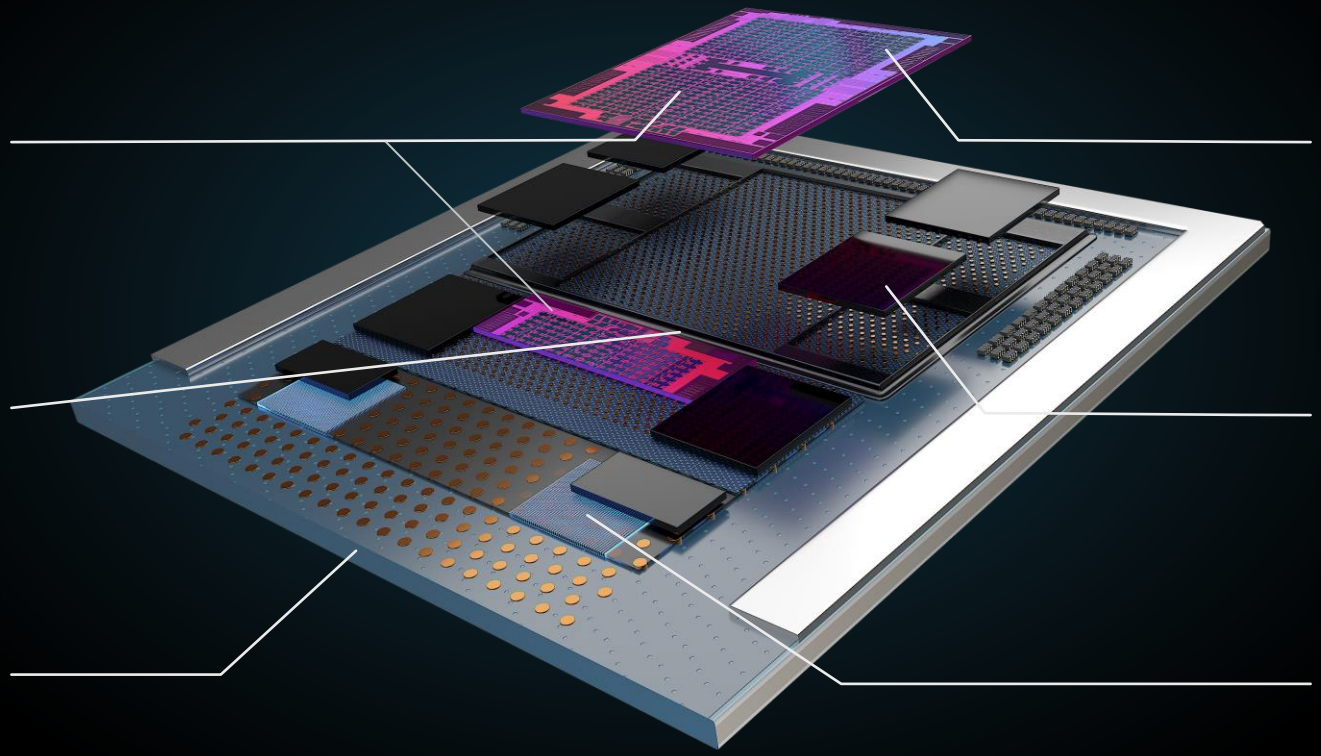
AMD INSTINCT™ MI200 SERIES

KEY INNOVATIONS

TWO AMD CDNA™2 CHIPLETS

ULTRA HIGH BANDWIDTH CHIPLET INTERCONNECT

COHERENT CPU-TO-GPU INTERCONNECT



2ND GEN MATRIX CORES FOR HPC & AI

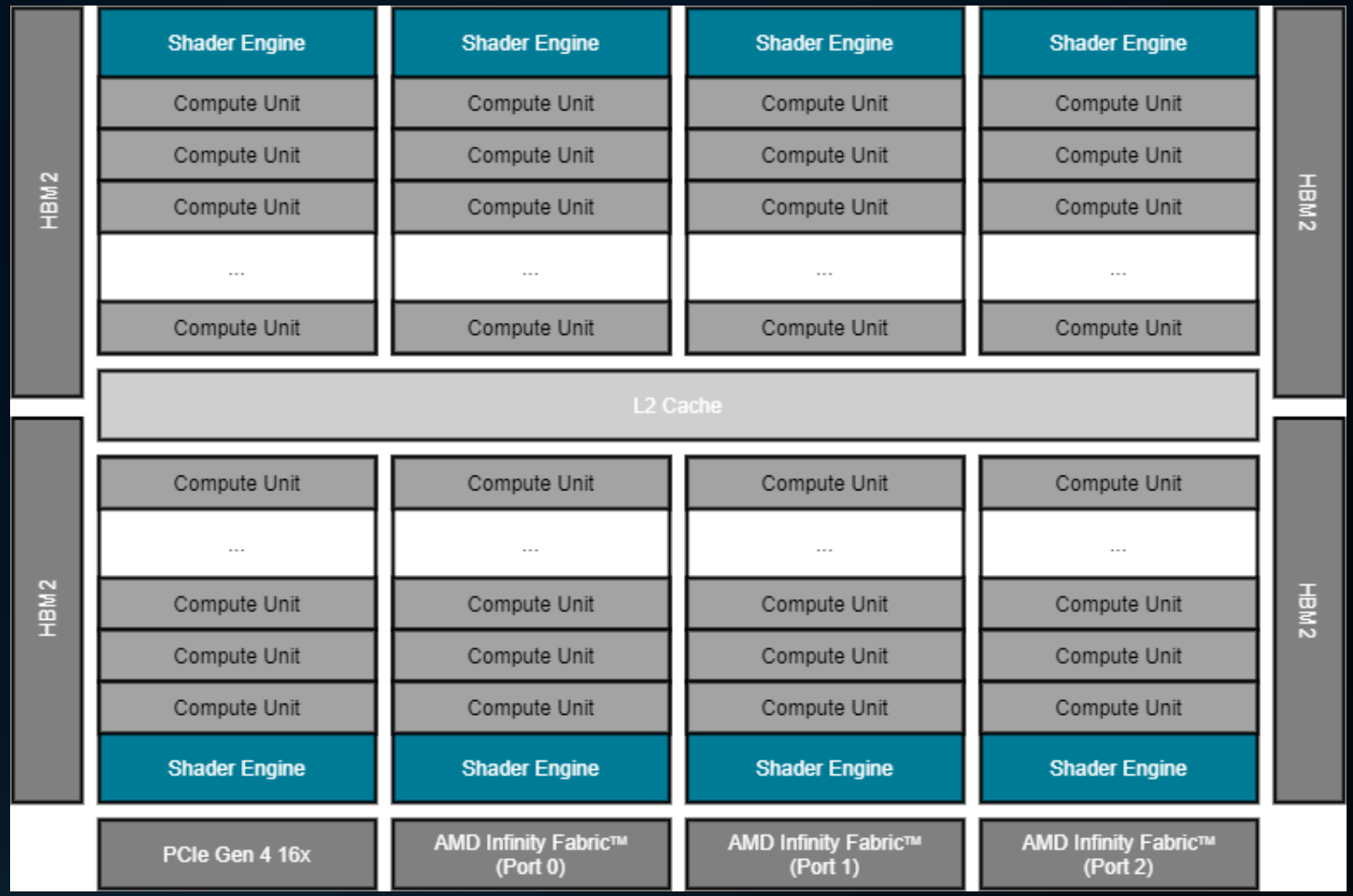
EIGHT STACKS OF HBM2E

2.5D ELEVATED FANOUT BRIDGE (EFB)

AMD INSTINCT™ MI200 OAM SERIES

AMD CDNA™ Architecture – Made for Performance

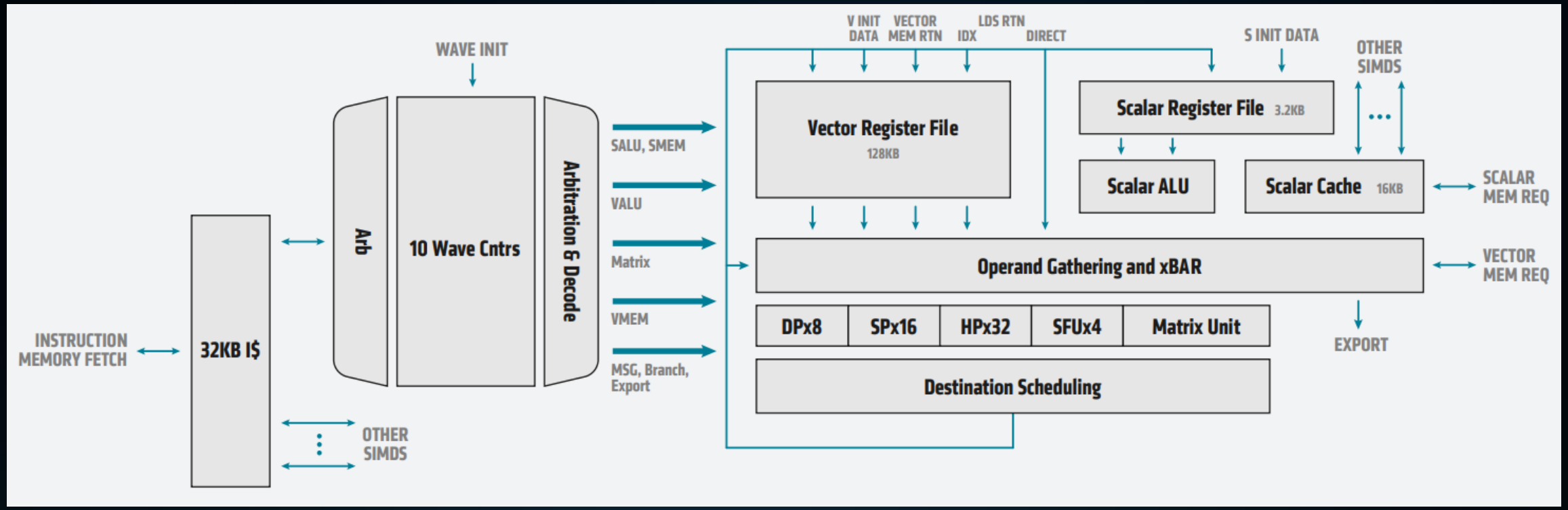
- GPU is composed from several main blocks using an on-die fabric
- 120 Compute Units (CU)
  - Four Compute Engines w/ SIMD
  - SIMD pipelines execute 16-wide instructions
- Support for int8, FP16, FP32, FP64, bfloat16



Compute Unit: Execution Units

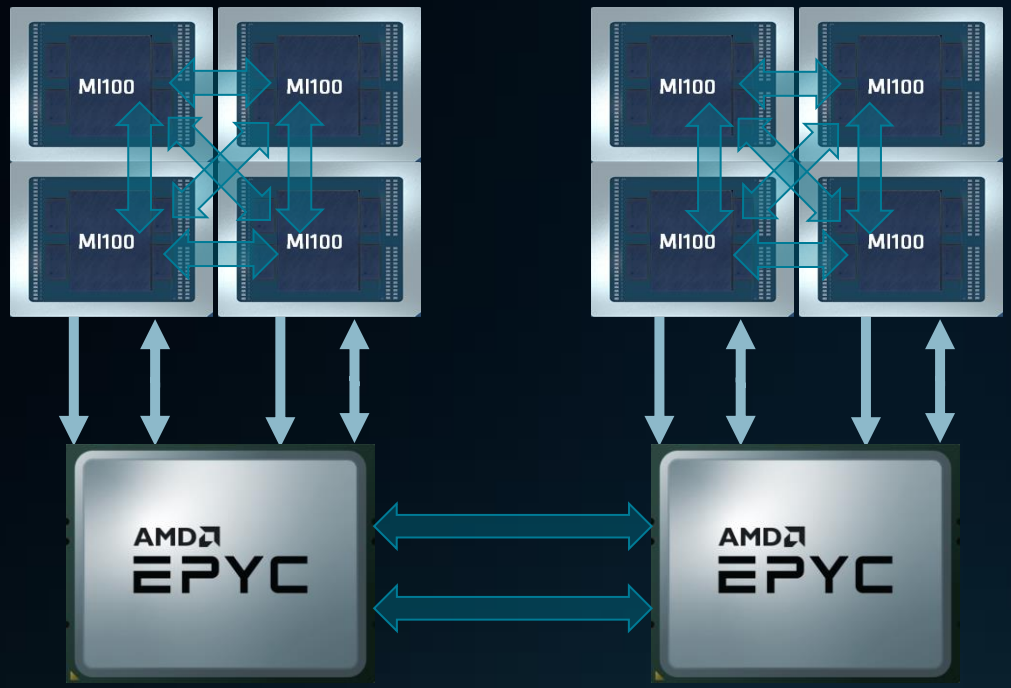
- Scheduling granularity: “wavefront” (64-wide)
- Split into 4x 16-wide SIMD operations
- Can schedule on SIMD instruction per cycle

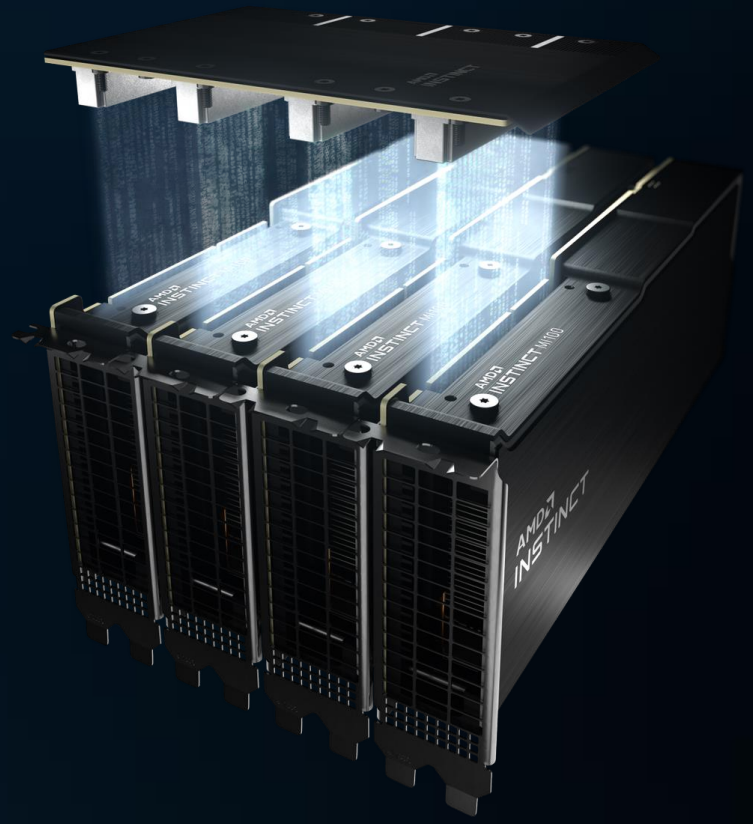
- Registers:
  - 256 VGPR (256 max. per wavefront)
  - 800 SGPR (102 max. per wavefront)



Node-level Design – GPU Hives

- GPUs can form “hives” of four GPUs
- One hive is associated to one processor
- High-speed AMD Infinity Fabric™ connections in the hives (fully connected).

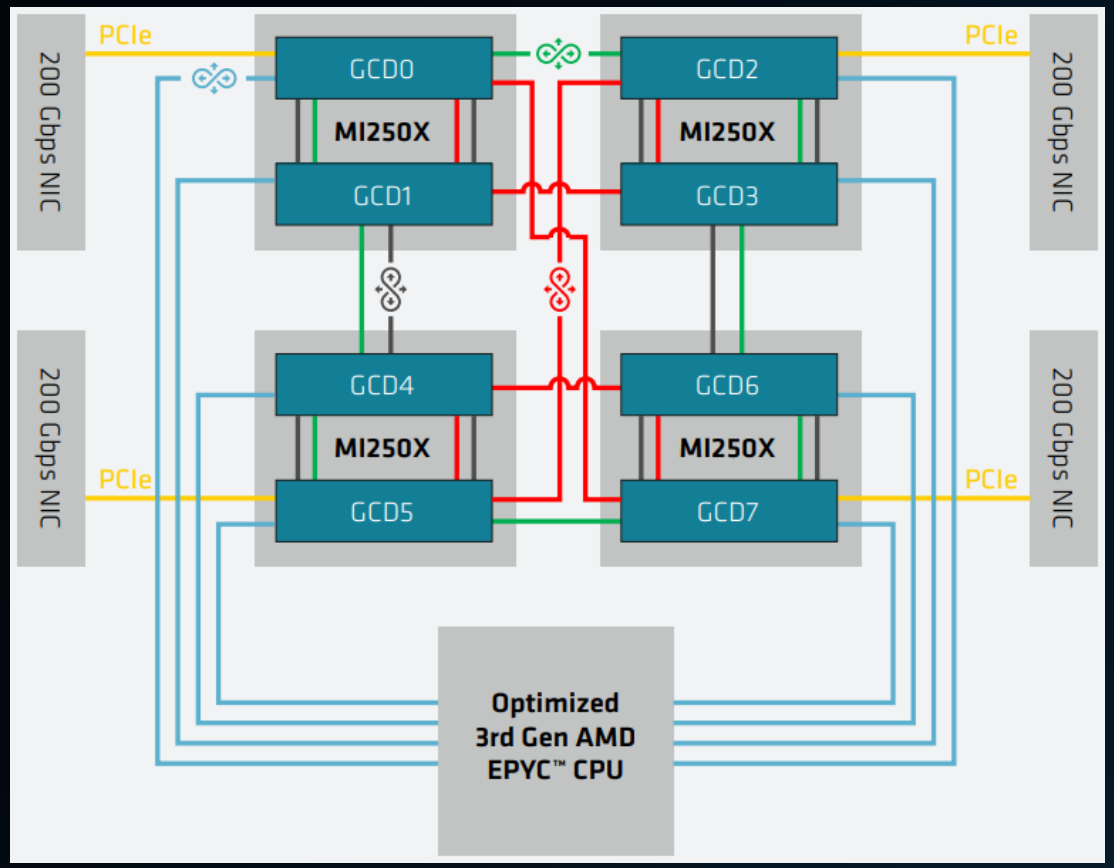




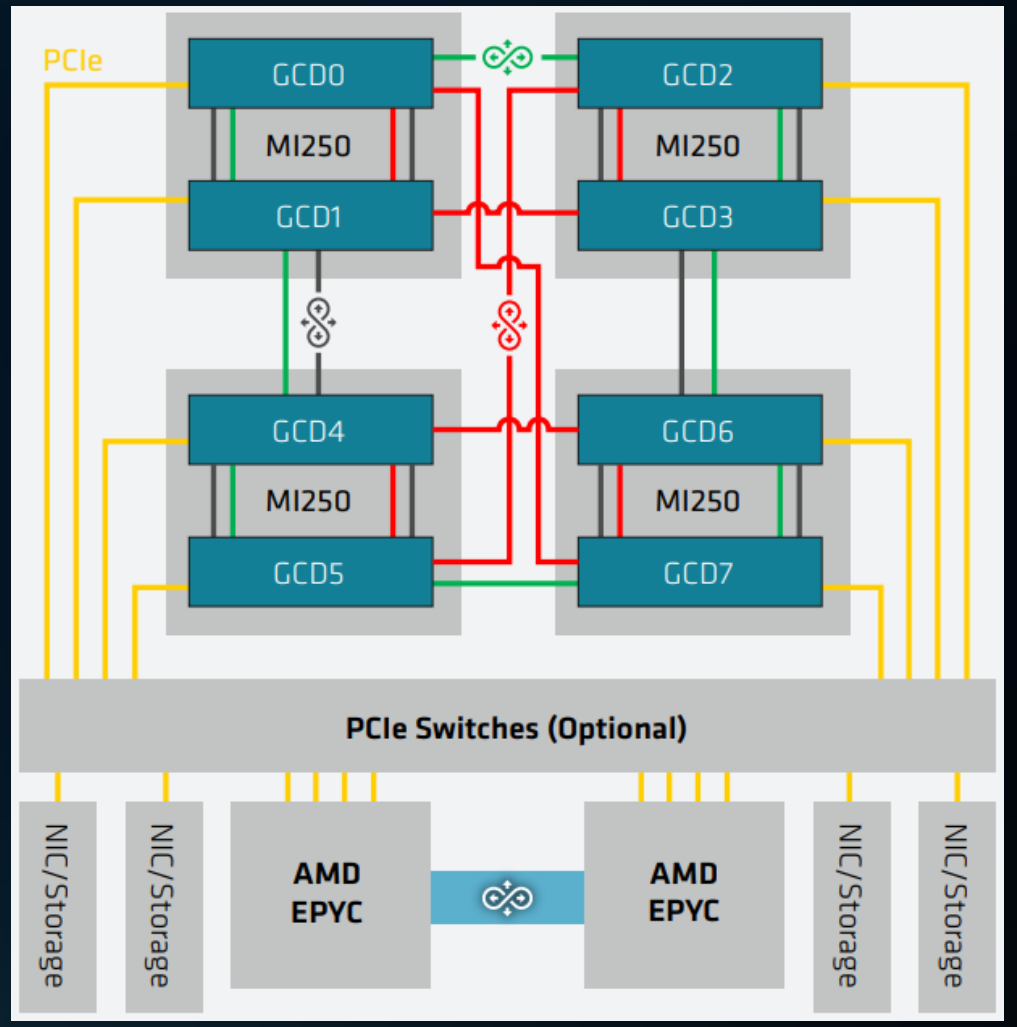
PCIe® Gen 4 link

AMD Infinity Fabric™ technology

AMD CDNA™ 2 Architecture: Node-level Design

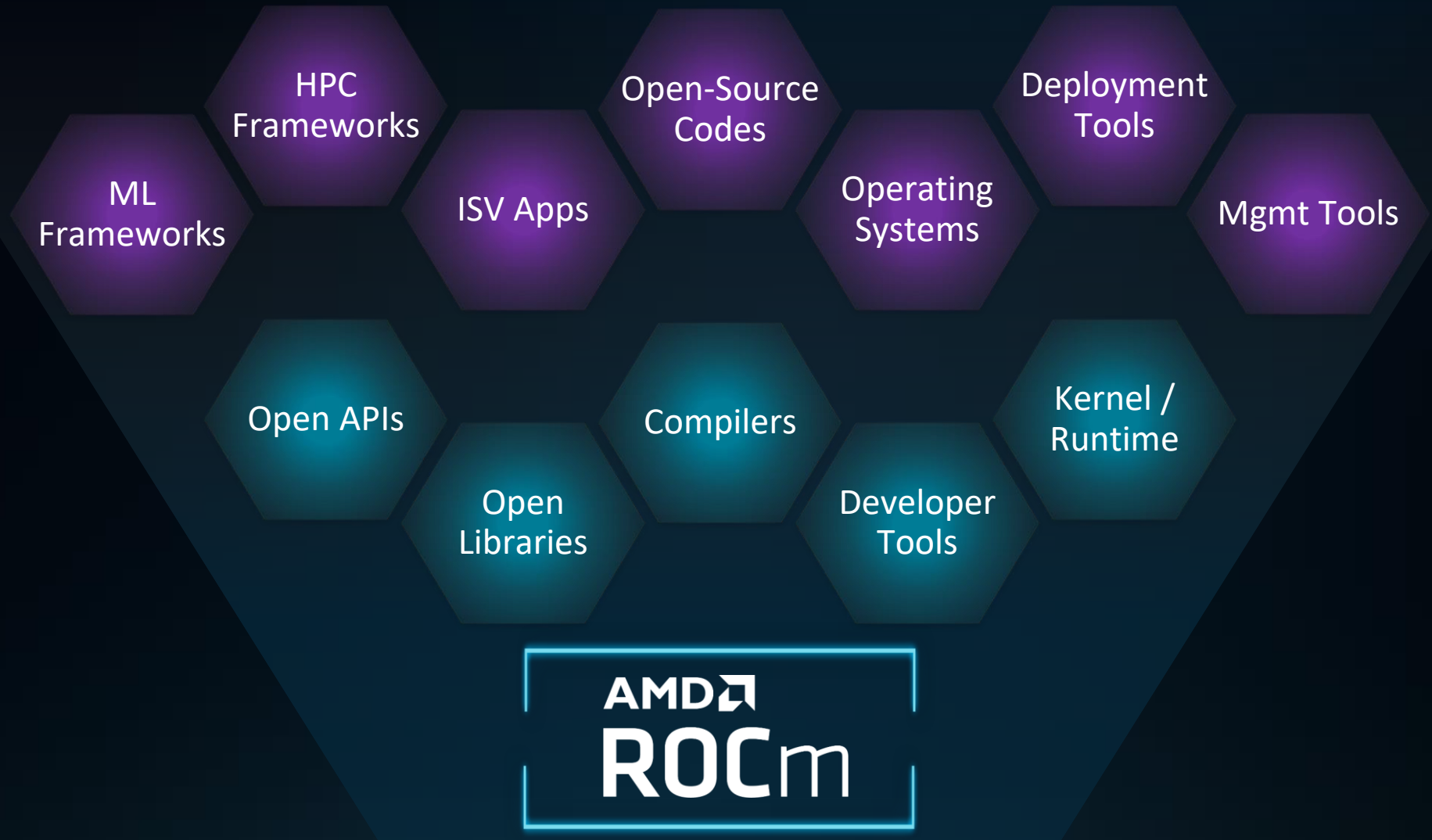


Green, Red, Gray, and Blue lines are AMD Infinity Fabric™ Links
Red and Green links can create two bi-directional rings
Blue Infinity Fabric Link provides coherent GCD-CPU connection



AMD ROCm™ Software Stack

ROCm™: Enabling An Ecosystem Without Borders

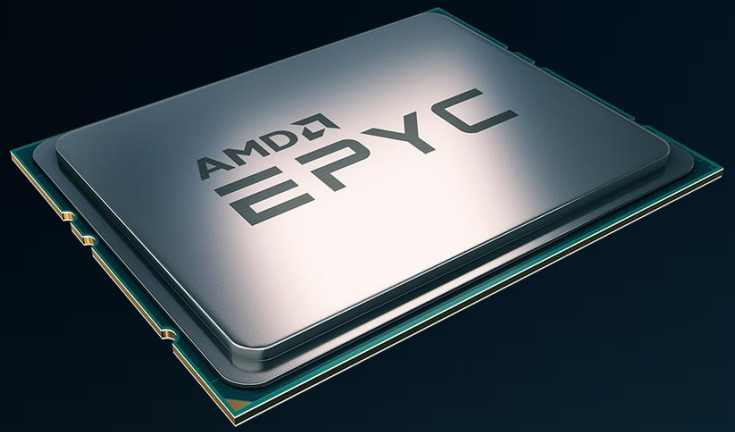


A Tale of Host and Device

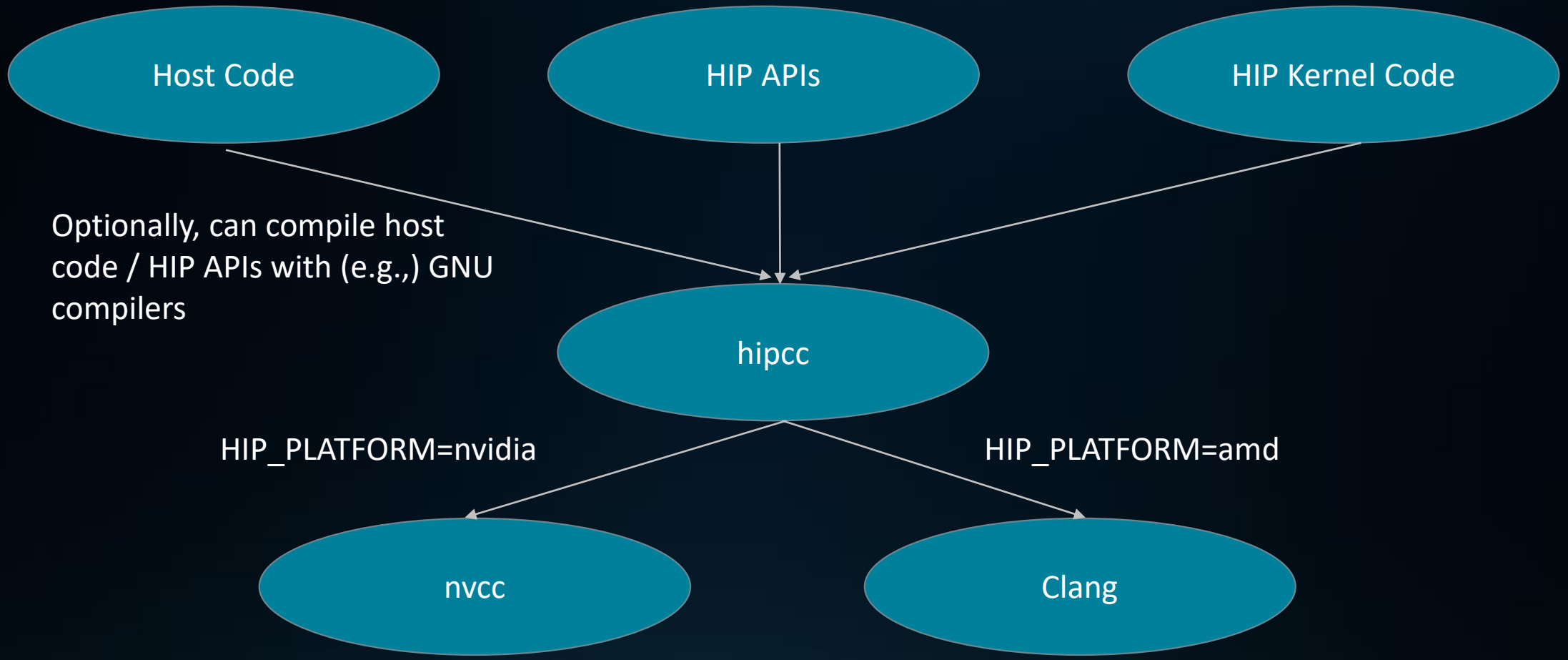
Source code in HIP has two flavors: Host code and Device code

- The host is the CPU.
- Host code runs here.
- Usual C++ syntax and features.
- Entry point is the ‘main’ function.
- HIP API can be used to create device buffers, move between host and device, and launch device code.

- The device is the GPU.
- Device code runs here.
- Device codes are launched via “kernels”
- Instructions from the Host are enqueued into “streams”.



Compiling your HIP code



Example: Calling BLAS Level 3 Routines (SGEMM)

Calling standard math library (host):

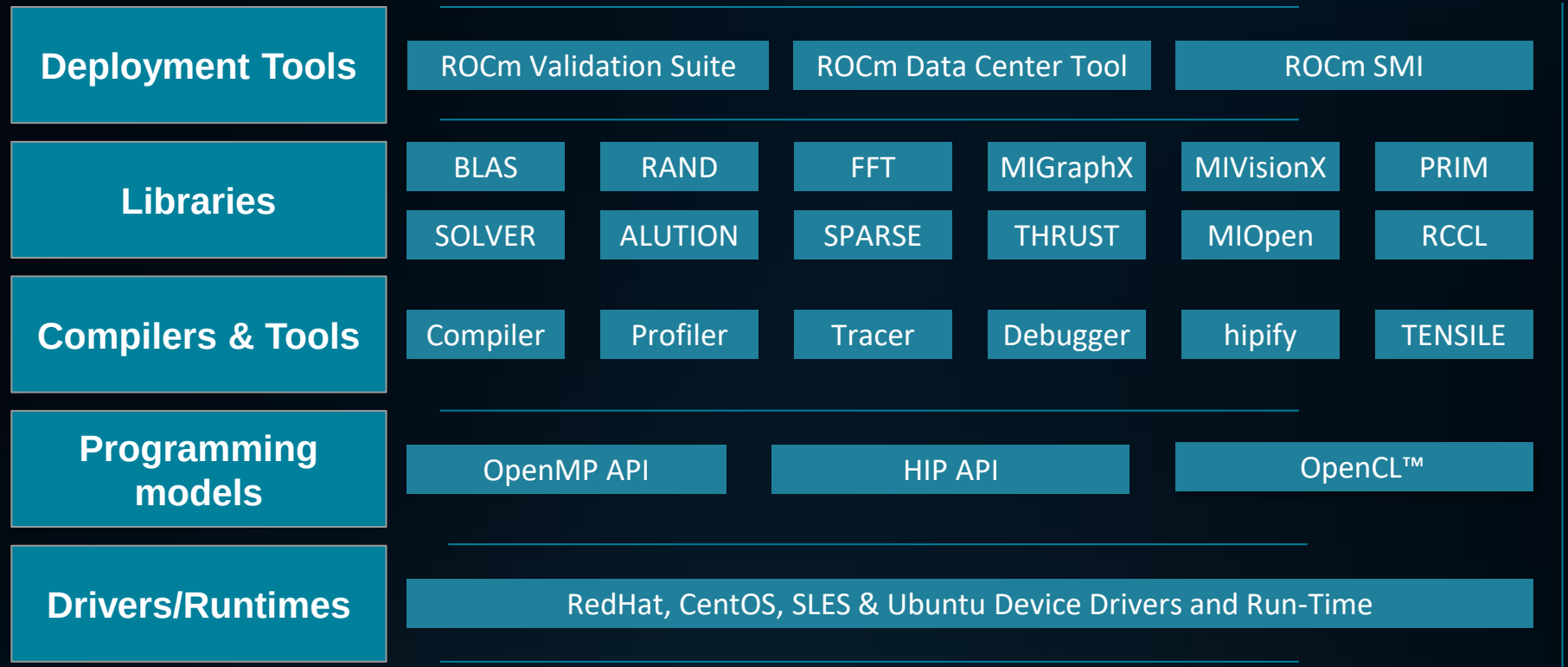
```
void example_sgemm_host() {
    // Declarations omitted.



    cblas_sgemm(transa, transb,
                m, n, k,
                alpha, A, lda,
                B, ldb,
                beta, C, ldc);


}
```

Calling rocBLAS math library (GPU):

```
void example_sgemm_gpu() {
    // Declarations omitted.
    // Assume matrix on GPU.
    rocblas_handle handle;
    rocblas_create_handle(&handle);
    rocblas_sgemm(handle,
                  transa, transb,
                  m, n, k,
                  &alpha, A, lda,
                  B, ldb,
                  &beta, C, ldc);
    rocblas_destroy_handle(handle);
}
```

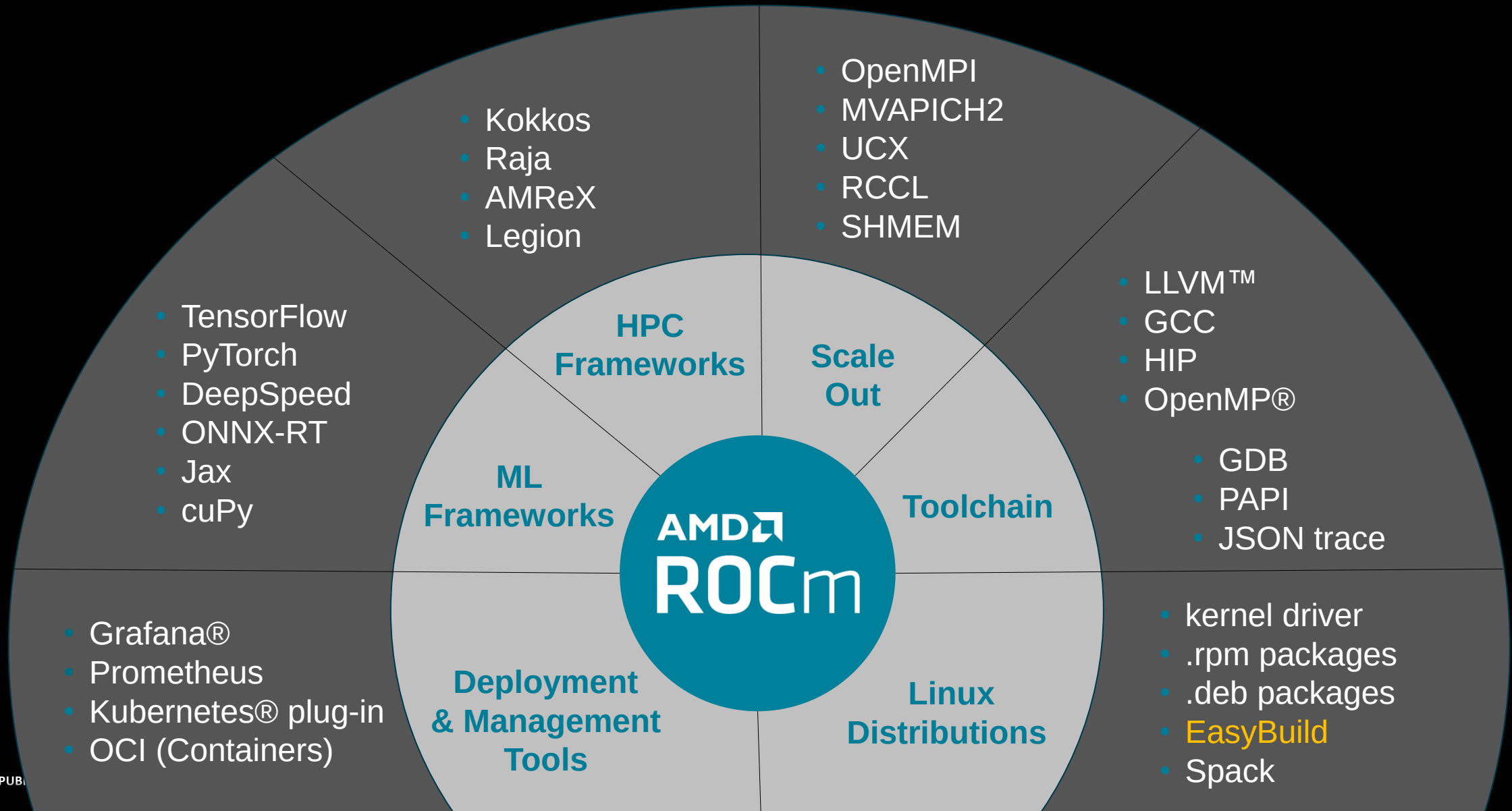
Library interface almost identical and easy to port from host usage to GPU usage.

AMD ROCm™ - The Core Components



Ecosystem

ROCM™ ENABLES THE ECOSYSTEM WITH SUPPORT OF OPEN APIS & TOOLS



Many EasyBuild Configurations Available Today

https://github.com/easybuilders/easybuild-easyconfigs/

Package	Description
/a/AOCC	AMD Optimized Compiler (EPYC)
/a/AOMP	OpenMP Compiler with target offload
/h/hipify-clang/	Conversion utility for CUDA to HIP
/r/ROCR-Runtime	Core ROCm component
/r/ROCT-Thunk-Interface	Abstraction layer for ROCm
/r/ROCm-CompilerSupport	GPU backend compiler
…	…

Special thanks to the contributors!

AMD INFINITY HUB

MORE APPS, MORE NUMBERS

AMD INSTINCT™ MI200 SUPPORT

Starting Now

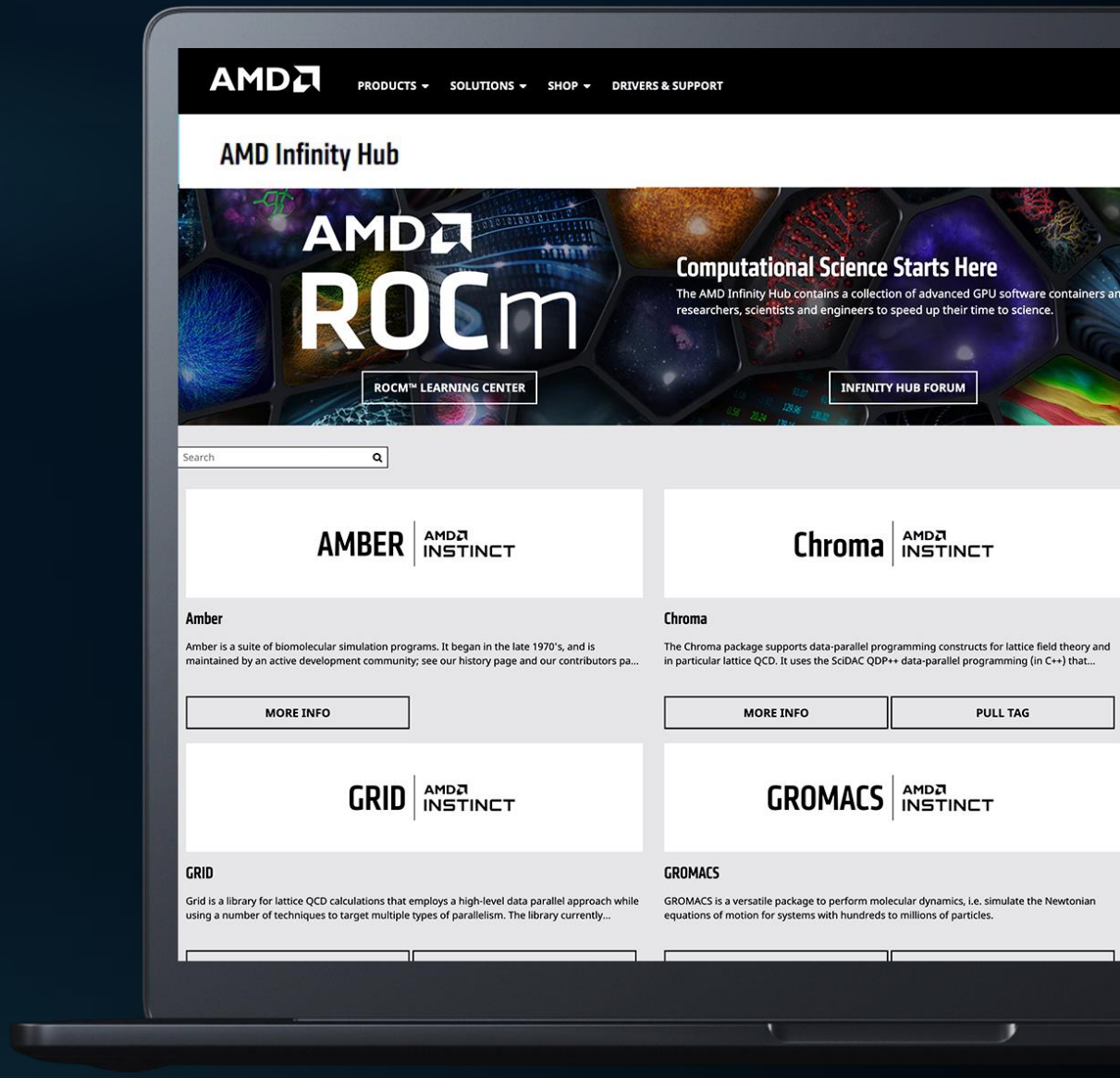
HPC APP GROWTH

Expanding with Weather, CFD, Quantum Chemistry and other codes

PERFORMANCE RESULTS

Published Performance Results for Select Apps / Benchmarks

AMD.com/InfinityHub



Getting Started with ROCm™ Open Software Platform

ROCm™ Learning Center

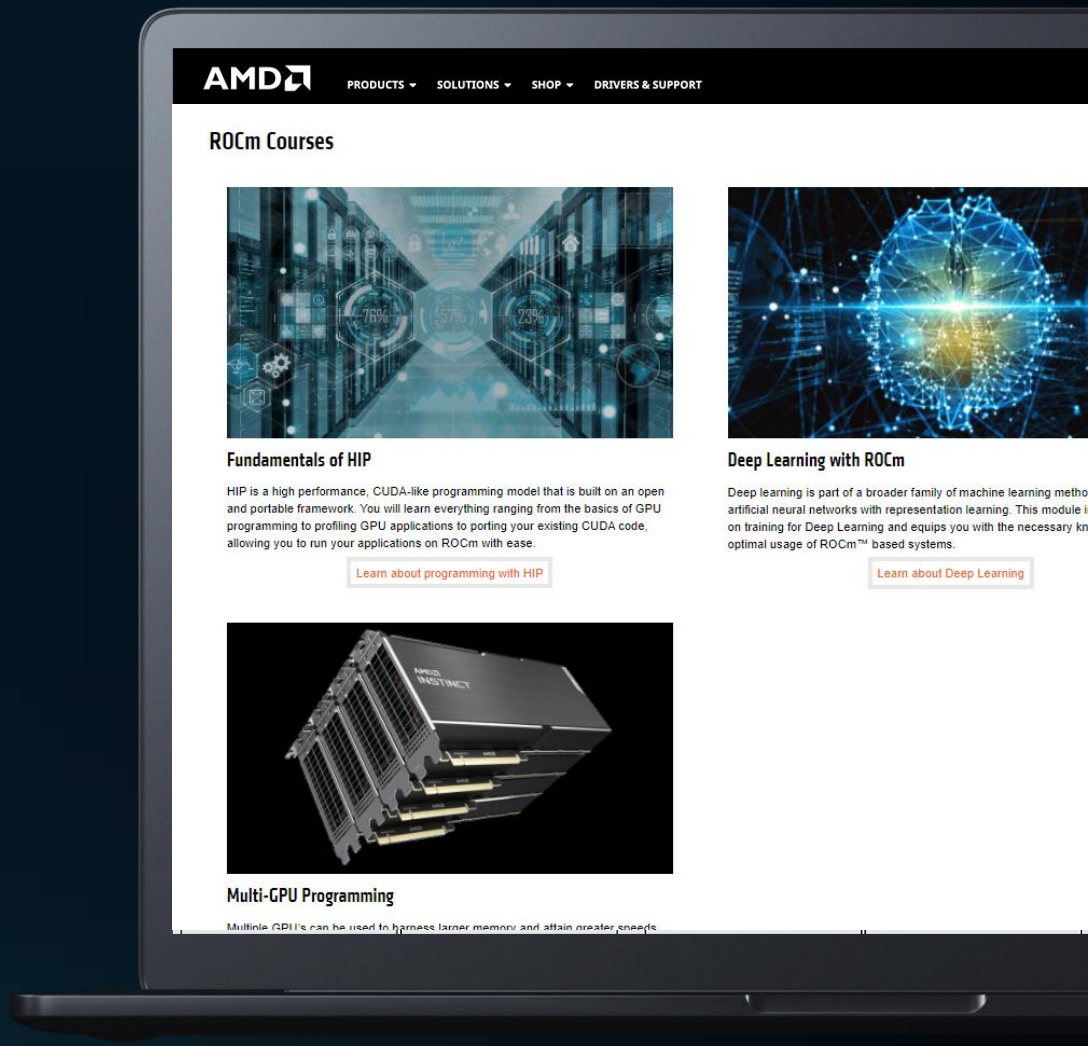
Curated videos, webinars, labs and tutorials for developers to learn how to use ROCm

developer.amd.com/resources/rocm-learning-center

AMD Accelerator Cloud

Remote access for customers and partners to test code and applications on the latest AMD GPUs

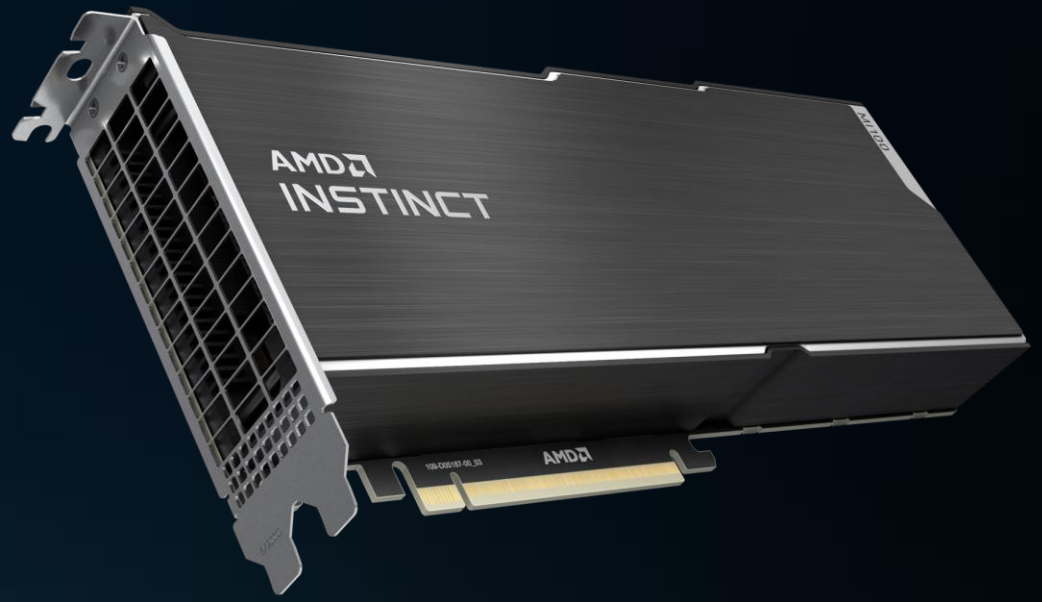
https://www.amd.com/en/solutions/accelerated-computing



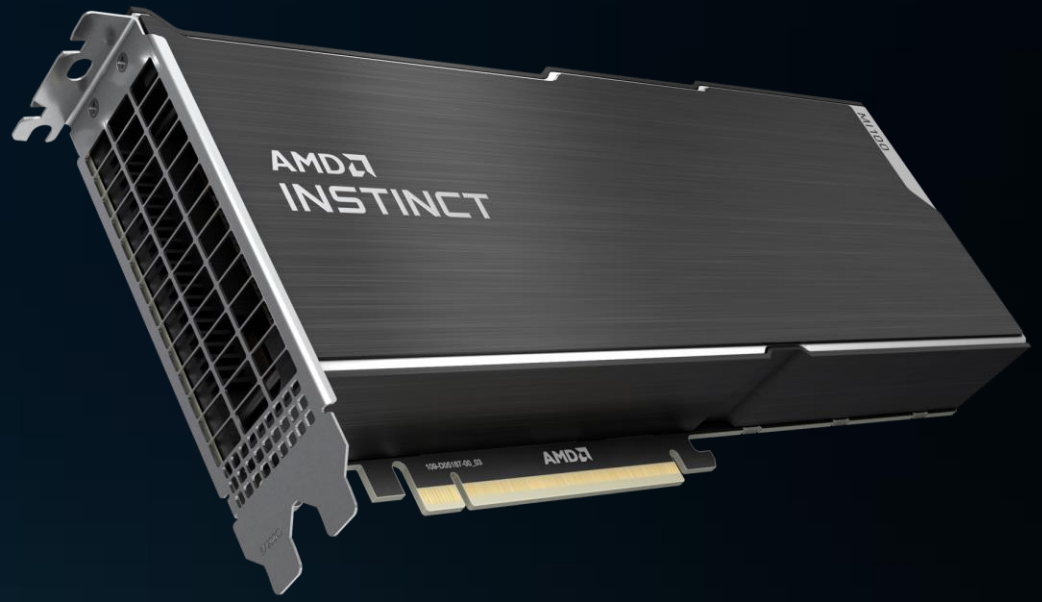
Summary

- AMD Instinct™ GPUs
  - High-performance GPU architecture designed for HPC and AI/ML

- AMD ROCm™ Software
  - Open-source!
  - Standards based: OpenMP
  - Portable: OpenMP, HIP
  - Easy to port: HIPification



Q&A



ENDNOTES

MI100-05

Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 11.535 TFLOPS peak theoretical double precision (FP64) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 6.62 TFLOPS FP64. Server manufacturers may vary configuration offerings yielding different results. MI100-05

MI100-14

Testing Conducted by AMD performance labs as of October 30th, 2020, on three platforms and software versions typical for the launch dates of the Radeon Instinct MI25 (2018), MI50 (2019) and AMD Instinct MI100 GPU (2020) running the benchmark application Quicksilver. MI100 platform (2020): Gigabyte G482-Z51-00 system comprised of Dual Socket AMD EPYC™ 7702 64-Core Processor, AMD Instinct™ MI100 GPU, ROCm™ 3.10 driver, 512GB DDR4, RHEL 8.2 MI50 platform (2019): Supermicro® SYS-4029GP-TRT2 system comprised of Dual Socket Intel Xeon® Gold® 6132, Radeon Instinct™ MI50 GPU, ROCm 2.10 driver, 256 GB DDR4, SLES15SP1 MI25 platform (2018): Supermicro SYS-4028GR-TR2 system comprised of Dual Socket Intel Xeon CPU E5-2690, Radeon Instinct™ MI25 GPU, ROCm 2.0.89 driver, 246GB DDR4 system memory, Ubuntu 16.04.5 LTS. MI100-14

MI100-15

Testing Conducted by AMD performance labs as of October 30th, 2020, on three platforms and software versions typical for the launch dates of the Radeon Instinct MI25 (2018), MI50 (2019) and AMD Instinct MI100 GPU (2020) running the benchmark application TensorFlow ResNet 50 FP 16 batch size 128. MI100 platform (2020): Gigabyte G482-Z51-00 system comprised of Dual Socket AMD EPYC™ 7702 64-Core Processor, AMD Instinct™ MI100 GPU, ROCm™ 3.10 driver, 512GB DDR4, RHEL 8.2 MI50 platform (2019): Supermicro® SYS-4029GP-TRT2 system comprised of Dual Socket Intel Xeon® Gold® 6254, Radeon Instinct™ MI50 GPU, ROCm 3.0.6 driver, 338 GB DDR4, Ubuntu® 16.04.6 LTS MI25 platform (2018): a Supermicro SYS-4028GR-TR2 system comprised of Dual Socket Intel Xeon CPU E5-2690, Radeon Instinct™ MI25 GPU, ROCm 2.0.89 driver, 246GB DDR4 system memory, Ubuntu 16.04.5 LTS. MI100-15

Disclaimer and Attributions

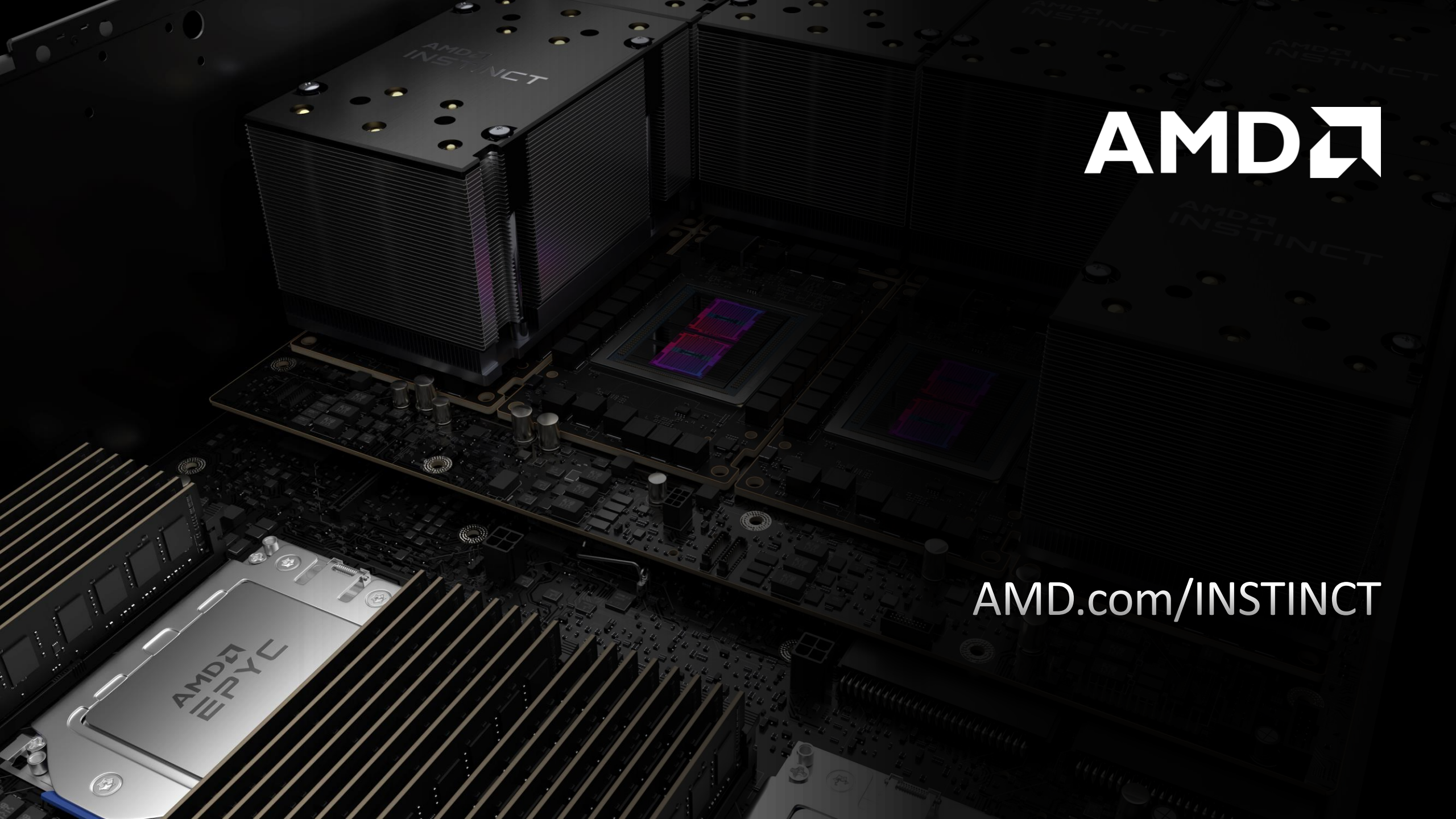
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.


AMD


AMD
AMD.com/INSTINCT

ENDNOTES (MI200-01 thru MI200-18)

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

MI200-02 - Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X accelerator (128GB HBM2e OAM module) at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak double precision matrix (FP64 Matrix) theoretical, floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator resulted in 19.5 TFLOPS peak double precision (FP64 Tensor Core) theoretical, floating-point performance. Results found at: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-02

MI200-07 - Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). The highest published results on the NVidia Ampere A100 (80GB) SXM GPU accelerator resulted in 2.039 TB/s GPU memory bandwidth performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf MI200-07

MI200-15 - Testing Conducted by AMD performance lab as of 10/7/2021, on a single socket AMD EPYC™ 'Trento' server, with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology, using LAMMPS ReaxFF/C, patch_2Jul2021 plus AMD optimizations to LAMMPS and Kokkos that are not yet available upstream resulted in a median score of 4x MI250X = 19,482,180.48 ATOM-Time Steps/s Vs. Dual AMD EPYC 7742@2.25GHz CPUs with 4x NVIDIA A100 SXM 80GB (400W) using LAMMPS classical molecular dynamics package ReaxFF/C, patch_10Feb2021 resulted in a published score of 8,850,000 (8.85E+06) ATOM-Time Steps/s. https://developer.nvidia.com/hpc-application-performance 19,482,180.48/8,850,000=2.20x (220%) the/1.2x (120%) faster. Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:lammps Information on LAMMPS: https://www.lammps.org/index.html Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-15

MI200-16 - Testing Conducted by AMD performance lab as of 10/18/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU powered server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology, using HACC, plus AMD optimizations to HACC that are not yet available upstream resulted in a median score of 1x MI250X = 4,400,000 (4.40E+06) Particles/s Vs. Testing Conducted by AMD performance lab as of 10/18/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W), using HACC resulted in a median score of 1x A100 = 2,290,000 (2.29E+06) Particles/s. Information on HACC: https://asc.llnl.gov/coral-2-benchmarks/gpu-versions-and-other-supplementary-material https://asc.llnl.gov/sites/asc/files/2020-09/coral-hacc-benchmark-summary-v1.7.pdf Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-16

MI200-17 - Testing conducted by AMD performance lab as of 10/13/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology, using LSMS, plus AMD optimizations to LSMS that are yet available upstream resulted in a median score of 1x MI250X = 3,950,000,000 (3.95E+09) Atom Interactions/s Vs. Testing conducted by AMD performance lab as of 9/27/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W), using LSMS resulted in a median score of 2,440,000,000 (2.44E+09) Atom Interactions/s. Information on LSMS: https://github.com/mstsuite/lsms, Information on GFortran: https://gcc.gnu.org/fortran/, Information on GCC Compiler: https://gcc.gnu.org/ Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-17

MI200-18 - Calculations conducted by AMD Performance Labs as of Sep 21, 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 HBMe memory capacity. Published specifications on the NVidia Ampere A100 (80GB) SXM and A100 accelerators (PCIe®) showed 80GB memory capacity. Results found at: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf MI200-18

ENDNOTES (MI200-19 thru MI200-25):

MI200-19 - Testing Conducted by AMD performance lab as of 10/1/2021, on a single socket AMD EPYC™ 'Trento' server with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology running AMG (Set up) FOM, resulting in a median score of 4x MI250X = 16,773,660,000 FOM_Setup / Sec (Setup Phase Time) Vs. Testing Conducted by AMD performance lab as of 10/1/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 4x NVIDIA A100 SXM 80GB (400W) running AMG (Set up) FOM, resulting in a median score of 4x A100 = 5,507,144,000 FOM_Setup / Sec (Setup Phase Time). Information on AMG_Setup: https://asc.llnl.gov/coral-2-benchmarks , https://asc.llnl.gov/sites/asc/files/2020-09/AMG_Summary_v1_7.pdf , Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-19

MI200-20 - Testing Conducted by AMD performance lab as of 10/1/2021, on a single socket AMD EPYC™ 'Trento' server, with 4x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology using AMG (Solve) FOM resulting in a median score of 4x MI250X = 73,318,380,000 FOM_Solve / Sec (Solve Phase Time) Vs. Testing Conducted by AMD performance lab as of 10/1/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 4x NVIDIA A100 SXM 80GB (400W), using AMG (Solve) FOM resulting in a median score of 4x A100 = 31,476,470,000 FOM_Solve / Sec (Solve Phase Time). Information on AMG_Solve: https://asc.llnl.gov/coral-2-benchmarks, https://asc.llnl.gov/sites/asc/files/2020-09/AMG_Summary_v1_7.pdf Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-20

MI200-21 - Testing Conducted by AMD performance lab as of 9/22/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using Nvidia Nbody 32 CUDA sample version 11.2.152 converted to HIP plus AMD optimizations to Nbody 32 that are not yet available upstream resulting in a median score of 2.3x MI250X = 31.72 Particles (Body-to-Body) Interactions/s Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using Nbody 32 sample code version 11.2.152 resulting in a median score of 14.12 Particles (Body-to-Body) Interactions/s. Information on Nbody 32: https://developer.download.nvidia.com/compute/DevZone/C/html_x64/Physically-Based_Simulation.html, https://github.com/AMD-HPC/nbody-nvidia. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-21

MI200-22 - Testing Conducted by AMD performance lab as of 9/22/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server with AMD Infinity Fabric™ technology, using Nbody 64 CUDA Sample version 11.2.152 converted to HIP. Nvidia Nbody 64 samples code version 11.2.152, plus AMD optimizations to Nbody 64 that are not yet available upstream resulted in a median score of 19.245 Particles (Body-to-Body) Interactions/s. Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark Nvidia Nbody 64 sample code version 11.2.152 resulting in a median score of 7.631 Particles (Body-to-Body) Interactions/s. Information on Nbody 64: https://developer.download.nvidia.com/compute/DevZone/C/html_x64/Physically-Based_Simulation.html, https://github.com/AMD-HPC/nbody-nvidia. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations. MI200-022

MI200-23 - Testing Conducted by AMD performance lab as of 10/6/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using Quicksilver - LLNL-CODE-684037 converted to HIP, plus AMD optimizations to Quicksilver that are on AMD Github branch resulted in a median score of 214,000,000 Segments/s Vs. Testing Conducted by AMD performance lab as of 9/22/2021, on Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using Quicksilver - LLNL-CODE-684037 run with CUDA code version 11.2.152 resulted in a median score of 85,500,000 Segments/s. Information on Quicksilver: AMD branch based on LLNL version for this testing: https://github.com/moes1/Quicksilver/tree/AMD-HIP , LLNL version: https://github.com/LLNL/Quicksilver & Quicksilver info sheet: https://hpc.llnl.gov/sites/default/files/Quicksilver_CTS.pdf. Note: A proxy app for the Monte Carlo Transport Code, Mercury. LLNL-CODE-684037. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-23

MI200-24 - Testing Conducted by AMD performance lab as of 10/12/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using benchmark OpenMM_amoebagk v7.6.0, (converted to HIP) and run at double precision (8 simulations*10,000 steps) plus AMD optimizations to OpenMM_amoebagk that are not yet upstream resulted in a median score of 387.0 seconds or 223.2558 NS/Day Vs. Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark OpenMM_amoebagk v7.6.0, run at double precision (8 simulations*10,000 steps) with CUDA code version 11.4 resulted in a median score of 921.0 seconds or 93.8111 NS/Day. Information on OpenMM: https://openmm.org/ Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-24

MI200-25 - Testing Conducted by AMD performance lab as of 9/30/2021, on a single socket AMD EPYC™ 'Trento' server with 1x AMD Instinctt™ MI250X OAM (128 GB HBM2e) 560W GPUs with AMD Infinity Fabric™ technology using MILC benchmark version 7.8.1 developer version MILC_QCD on Github, Apex Medium test module, plus AMD optimizations to MILC that are not yet available upstream resulted in a median score 1,604.567 Total Time (Seconds). Vs. Dual AMD EPYC 7742@2.25GHz CPUs with 1x NVIDIA A100 SXM 80GB (400W) using MILC benchmark version develop_c30ed15e (quda0.8-patch4Oct2017), Apex Medium test module, resulted in a published score of 2,262 Total Time (Seconds). https://developer.nvidia.com/hpc-application-performance Nvidia MILC Container details found at: https://ngc.nvidia.com/catalog/containers/hpc:milc Information on MILC: https://web.physics.utah.edu/~detar/milc/ MILC Manual Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-25

ENDNOTES (MI200-26 thru MI200-31 and MI100-03 thru MI100-04):

MI200-26 - Testing Conducted by AMD performance lab as of 10/14/2021, on a single socket 3rd Gen AMD EPYC™ 'Trento' CPU server, with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using benchmark HPL v2.3, plus AMD optimizations to HPL that are not yet upstream resulted in a median score of 42.26 TFLOPS Vs. Nvidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark HPL Nvidia container image 21.4-HPL resulting in a median score of 15.33 TFLOPS. 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 MI200-26

MI200-27 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

MI200-31 - As of October 20th, 2021, the AMD Instinct™ MI200 series accelerators are the "Most advanced server accelerators (GPUs) for data center," defined as the only server accelerators to use the advanced 6nm manufacturing technology on a server. AMD on 6nm for AMD Instinct MI200 series server accelerators. Nvidia on 7nm for Nvidia Ampere A100 GPU. https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/ MI200-31

Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps*(4,096 bits*2))/8). Calculations by AMD Performance Labs as of OCT 5th, 2020 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI100 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) MI200-33

MI100-03 - 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 double precision (FP64), 46.1 TFLOPS peak single precision matrix (FP32), 23.1 TFLOPS peak single precision (FP32), 184.6 TFLOPS peak half precision (FP16) peak theoretical, floating-point performance. Published results on the NVidia Ampere A100 (40GB) GPU accelerator resulted in 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16) theoretical, floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-03

MI100-04 - Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 184.57 TFLOPS peak theoretical half precision (FP16) and 46.14 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 26.5 TFLOPS peak theoretical half precision (FP16) and 13.25 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-04

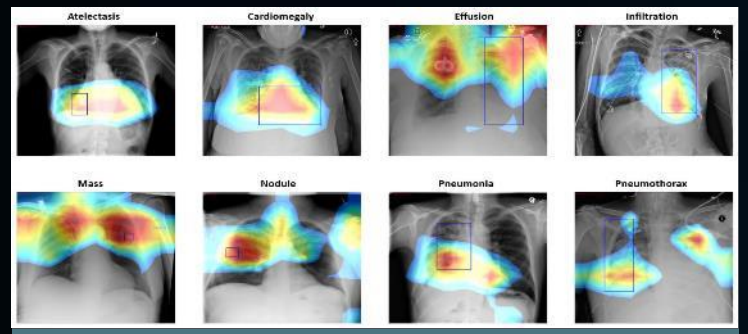
Backup

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	Planned Q1-2022	Docker Hub	Est 2022
CuPy	cupy.dev	Docker Hub	cupy.dev

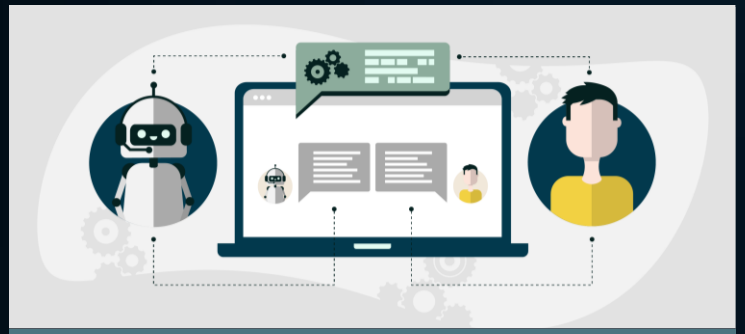
ML Models Supported on AMD ROCm™ Today



VIDEO & IMAGE RECOGNITION

Optimized Models
Resnet, VGG, Inception
GoogleNet, ResNext

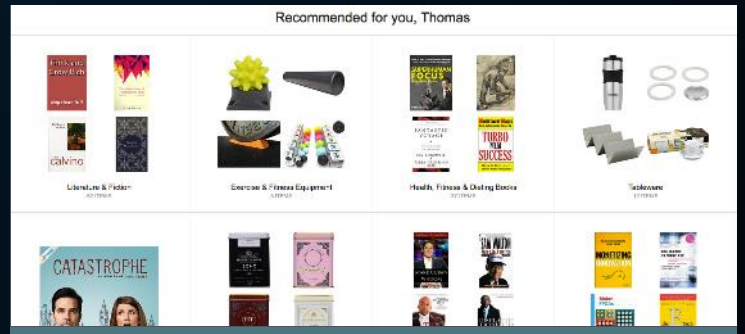
Markets
Automotive/Self Driving Cars
Healthcare/Medical Imaging
Public Safety



LANGUAGE PROCESSING

Optimized Models
GNMT, BERT, GPT-2

Markets
Customer Service
Web Services/E-Commerce



RECOMMENDATION ENGINE

Optimized Models
DLRM

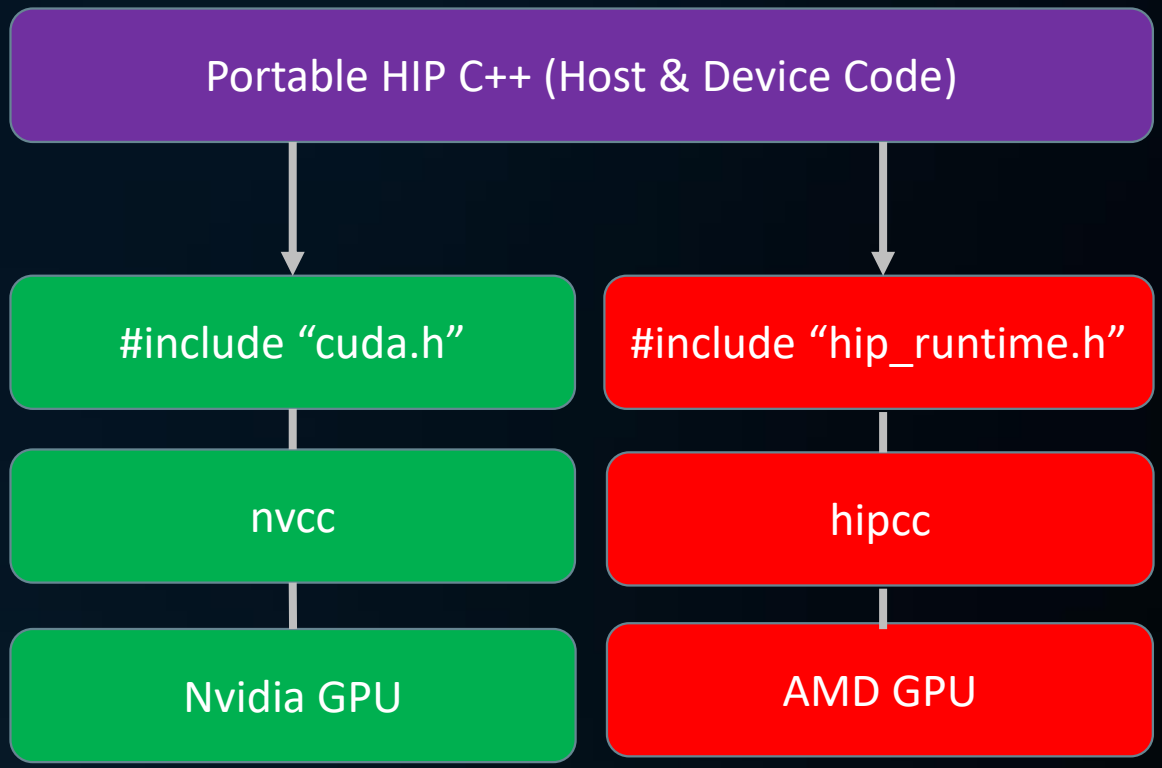
Markets
Web Services/E-commerce
SaaS

What is HIP?

AMD **H**eterogeneous-compute **I**nterface for **P**ortability, or **HIP**, is a C++ runtime API and kernel language that allows developers to create portable applications that can run on AMD's accelerators as well as CUDA devices.

HIP:

- Is open-source!
- Provides an API for an application to leverage GPU acceleration for the hardware of your choice.
- Syntactically similar to the CUDA® API enabling developers familiar with CUDA programming to easily extend their knowledge to new hardware platforms.
- Most CUDA API calls can be converted in place.
- Supports a strong subset of CUDA runtime functionality and enables creative developers to innovate on multiple hardware platforms.



Example: saxpy() – Very Common Operation in HPC Codes

```
void saxpy(size_t n, float a,
           float * x, float * y) {
    double t = 0.0;
    double tb, te;
    tb = omp_get_wtime();
    #pragma omp parallel for firstprivate(a)
    for (int i = 0; i < n; i++) {
        y[i] = a * x[i] + y[i];
    }
    te = omp_get_wtime();
    t = te - tb;
    printf("Time of kernel: %lf\n", t);
}
```

Timing code (not needed, just to have a bit more code to show ☺)

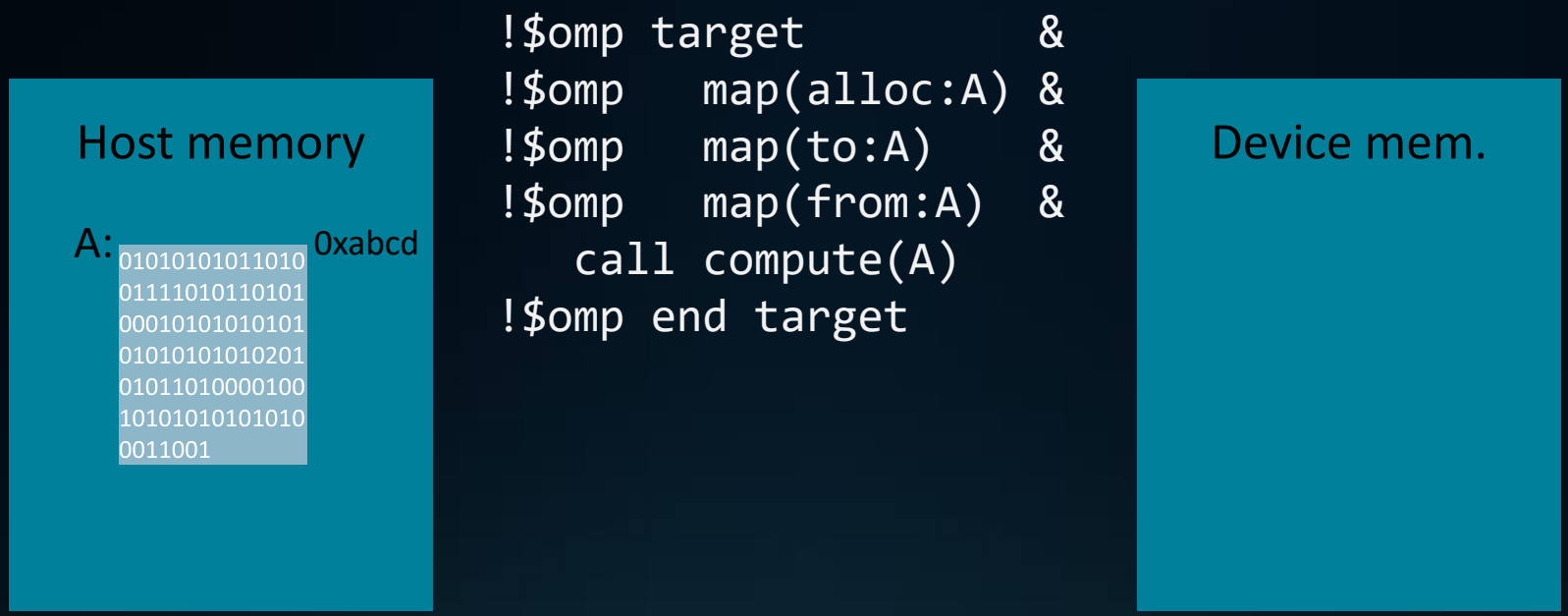
This is the code we want to execute on a target device (i.e., GPU)

Timing code (not needed, just to have a bit more code to show ☺)

Don’t do this at home!
Use a math library for this!

OpenMP: Heterogenous Programming (aka Offloading)

- As of version 4.0, the OpenMP API supports offloading computation to GPUs.
- Similar device model compared to other heterogenous programming models:
  - One host for “traditional” multi-threading.
  - Multiple GPUs of the same kind for offloading.
  - GPU devices are accessible though a device ID (from 0 to *n*-1 for *n* devices).



Example: saxpy() on a GPU

```
void saxpy(size_t n, float a,
           float * x, float * y) {
    double t = 0.0;
    double tb, te;
    tb = omp_get_wtime();
    #pragma omp target \
                teams distribute parallel for \
                map(to:x[0:SZ]) map(tofrom:y[0:SZ])
    for (int i = 0; i < SZ; i++) {
        y[i] = a * x[i] + y[i];
    }
    te = omp_get_wtime();
    t = te - tb;
    printf("Time of kernel: %lf\n", t);
}
```

- No need for boilerplate code to
  - allocate memory,
  - transfer data, and
  - synchronize GPU execution.
- Tightly integrates with multi-threaded execution on the host
- Directive-based language
  - Fortran!
  - (No need to switch to a different base language.)
- Descriptive and prescriptive model

HIP API

- Device Management:
  - `hipSetDevice()`, `hipGetDevice()`, `hipGetDeviceProperties()`
- Memory Management
  - `hipMalloc()`, `hipMemcpy()`, `hipMemcpyAsync()`, `hipFree()`, `hipHostMalloc()`
- Streams
  - `hipStreamCreate()`, `hipSynchronize()`, `hipStreamSynchronize()`, `hipStreamFree()`
- Events
  - `hipEventCreate()`, `hipEventRecord()`, `hipStreamWaitEvent()`, `hipEventElapsedTime()`
- Device Kernels
  - `__global__`, `__device__`, `hipLaunchKernelGGL()`
- Device code
  - `threadIdx, blockIdx, blockDim, __shared__`
  - 200+ math functions covering entire CUDA math library.
- Error handling
  - `hipGetLastError()`, `hipGetErrorString()`

HIP Kernel for saxpy()

```
__global__ void saxpy_kernel(size_t n, float a, float * x, float * y) {
    size_t i = threadIdx.x + blockIdx.x * blockDim.x;
    y[i] = a * x[i] + y[i];
}

void saxpy(size_t n, float a, float * x, float * y) {
    assert(n % 256 == 0);
    saxpy_kernel<<<n/256,256,0,NULL>>>(n, a, x, y);
}
```

HIPify Tools

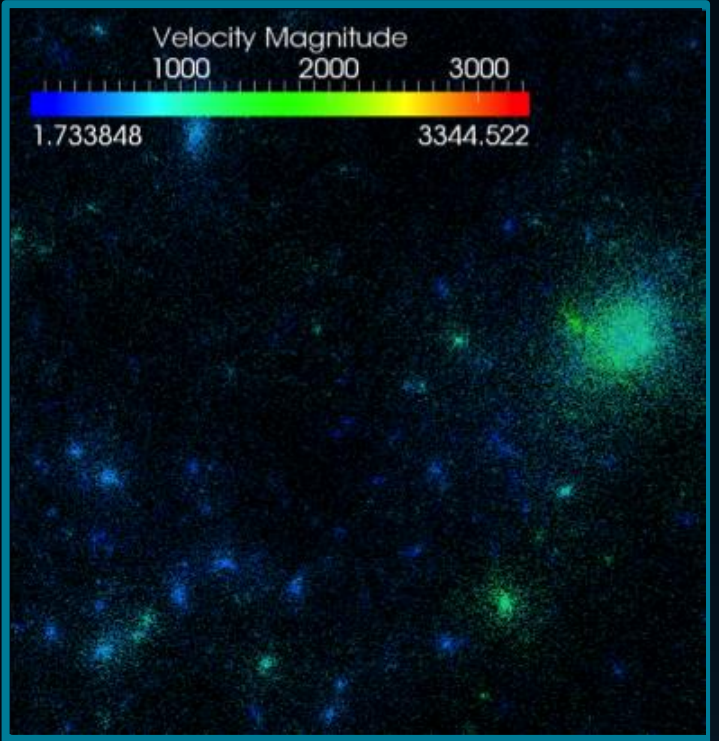
- The AMD ROCm™ platform provides ‘HIPification’ tools to do the heavy-lifting when porting CUDA code to the ROCm platform
  - hipify-perl
  - hipify-clang

- hipify-perl:
  - Easy to use – point at a directory and it will attempt to hipify CUDA code
  - Very simple string replacement technique: may make incorrect translations
    - `sed -e ‘s/cuda/hip/g’` (e.g., cudaMemcpy becomes hipMemcpy)
  - Recommended for quick scans of projects

- hipify-clang:
  - Requires the Clang compiler
  - More robust translation of the CUDA® API code
  - Uses clang to parse files and perform semantic translation
  - Can generate warnings and assistance for code for additional user analysis
  - High quality translation, particularly for cases where the user is familiar with the make system

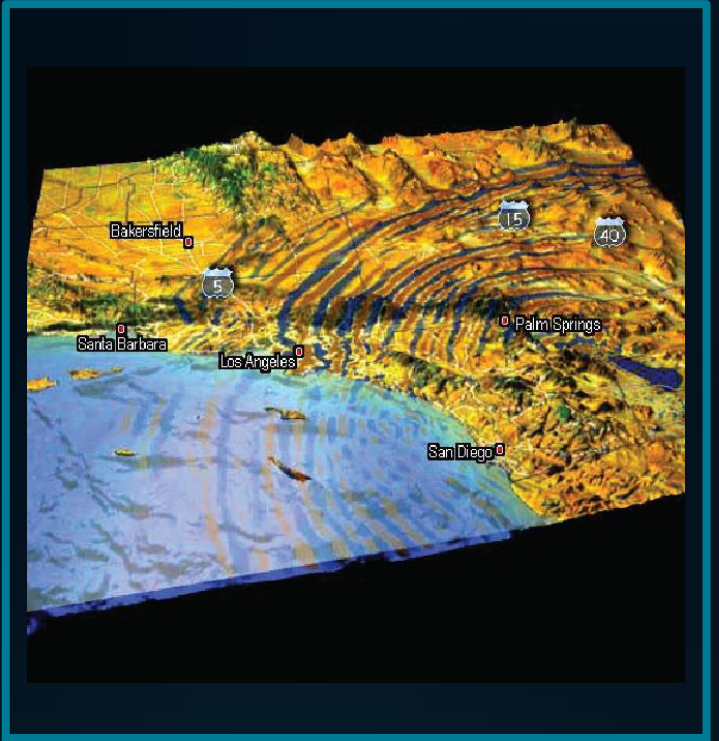
Seamlessly Porting CUDA Apps

Draw from the benefits of maintaining a single source that runs on multiple platforms



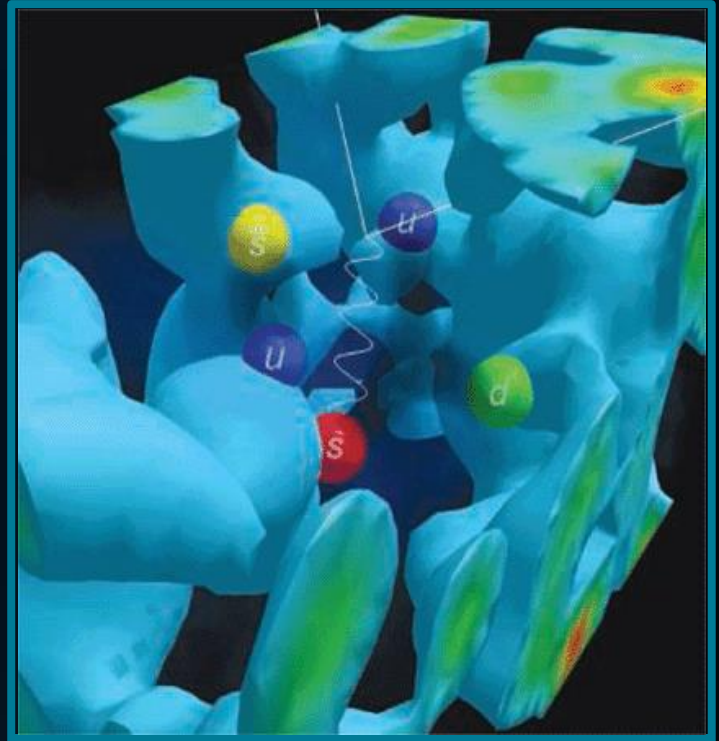
HACC

Ported in an Afternoon



SPECFEM3D

15K lines of CUDA, Ported in 1 Day



QUDA

Ported in 3 Weeks