AMD Instinct GPU and ROCm Technical Overview

PC Cluster Consortium 2022

Greg Oakes
AMD Senior HPC and AI Specialist

CAUTIONARY STATEMENT

This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) such as the features, functionality, performance, availability, timing and expected benefits of AMD products and AMD product roadmaps, 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.

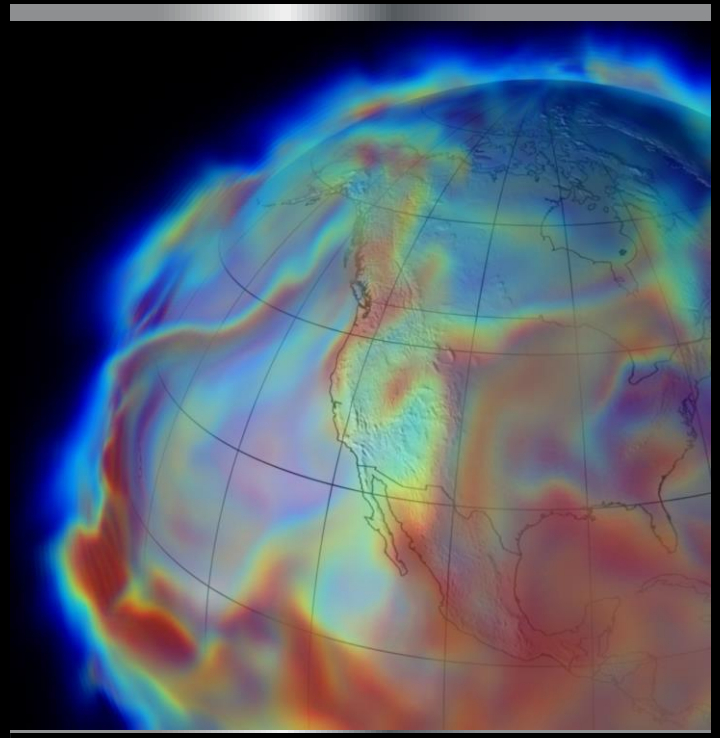
Agenda

- AMD Instinct GPU
- AMD ROCm Software Development Environment
- Q &A

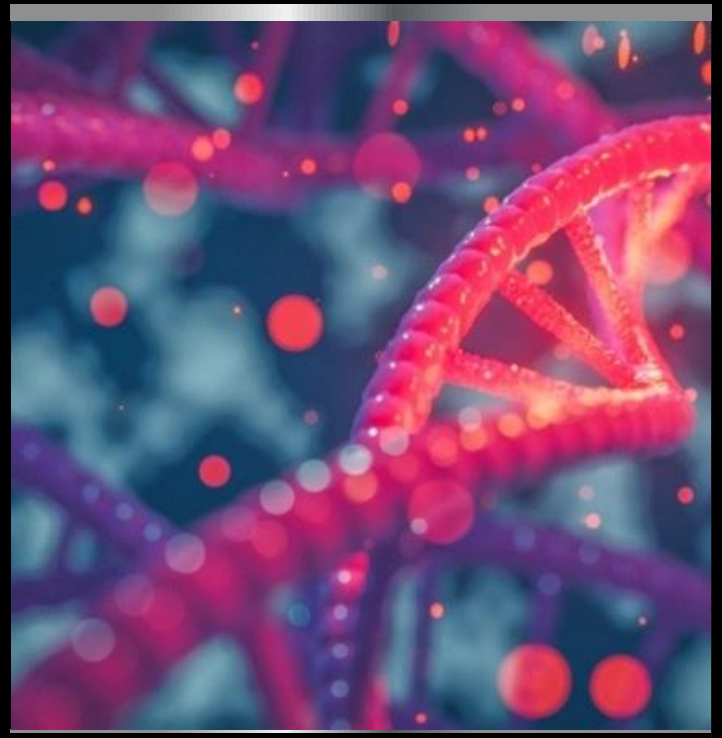
AMD

AMD Instinct™ CDNA2 Accelerators

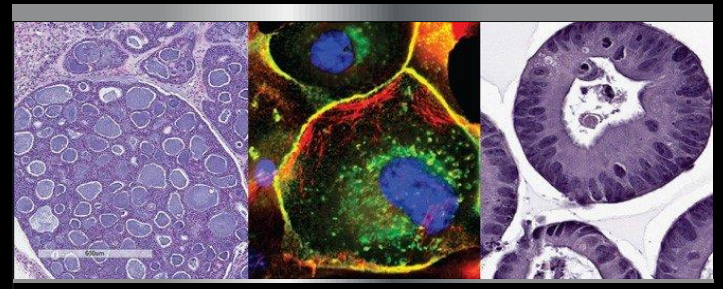
COMPUTATIONAL SCIENCE HAS NEVER BEEN MORE CRITICAL



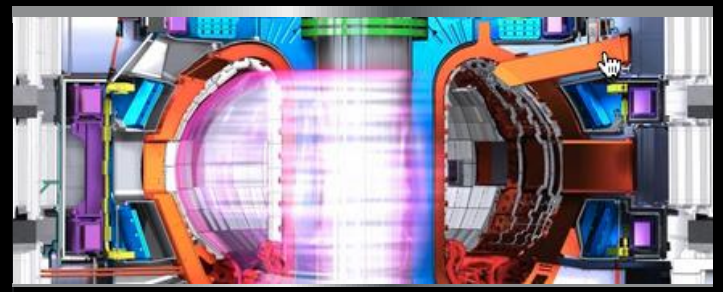
Climate Change
Simulating 50 Years into the Future



Understanding COVID Long Haulers
Genetic Analysis to Model Symptoms & Effects



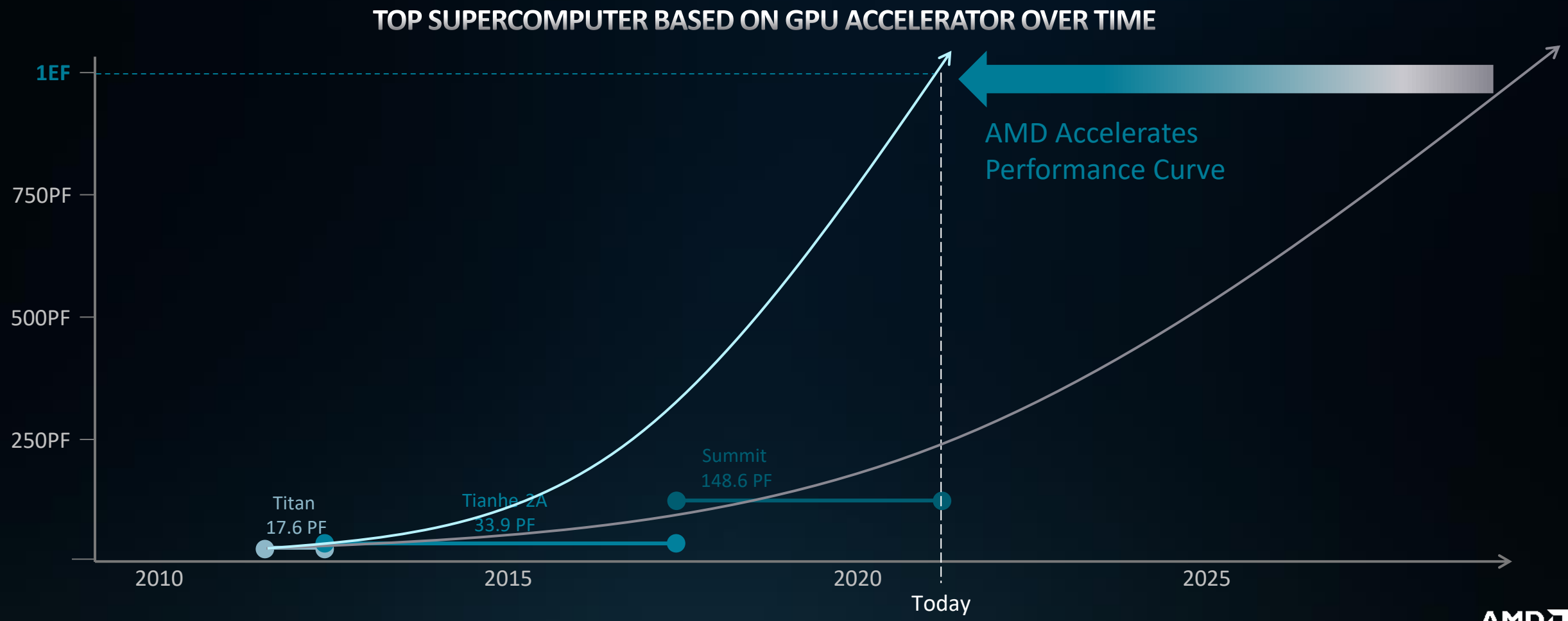
Fighting Cancer
Using AI to Develop Drug Therapy



Clean Energy
Building Safe Fusion Reactor

NEW ERA HAS ARRIVED- YEARS EARLIER

ENABLING SCIENTISTS TO TAKE GREAT COMPUTATIONAL CHALLENGES HEAD ON



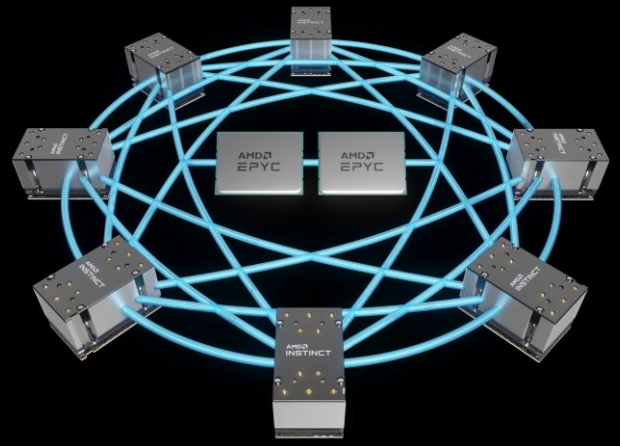
AMD

AMD Platform for Accelerated Computing

LEADERSHIP PERFORMANCE FOR HPC & AI

AMD CDNA 2

WORKLOAD-OPTIMIZED COMPUTE ARCHITECTURE



3RD GEN AMD INFINITY ARCHITECTURE

AMD ROCm

OPEN & PORTABLE SOFTWARE

AMD INSTINCT™ MI200 SERIES

AMD INSTINCT™

MI200 OAM

MI250, MI250X

AMD INSTINCT™

MI210 PCIe®

COMING SOON

AMD

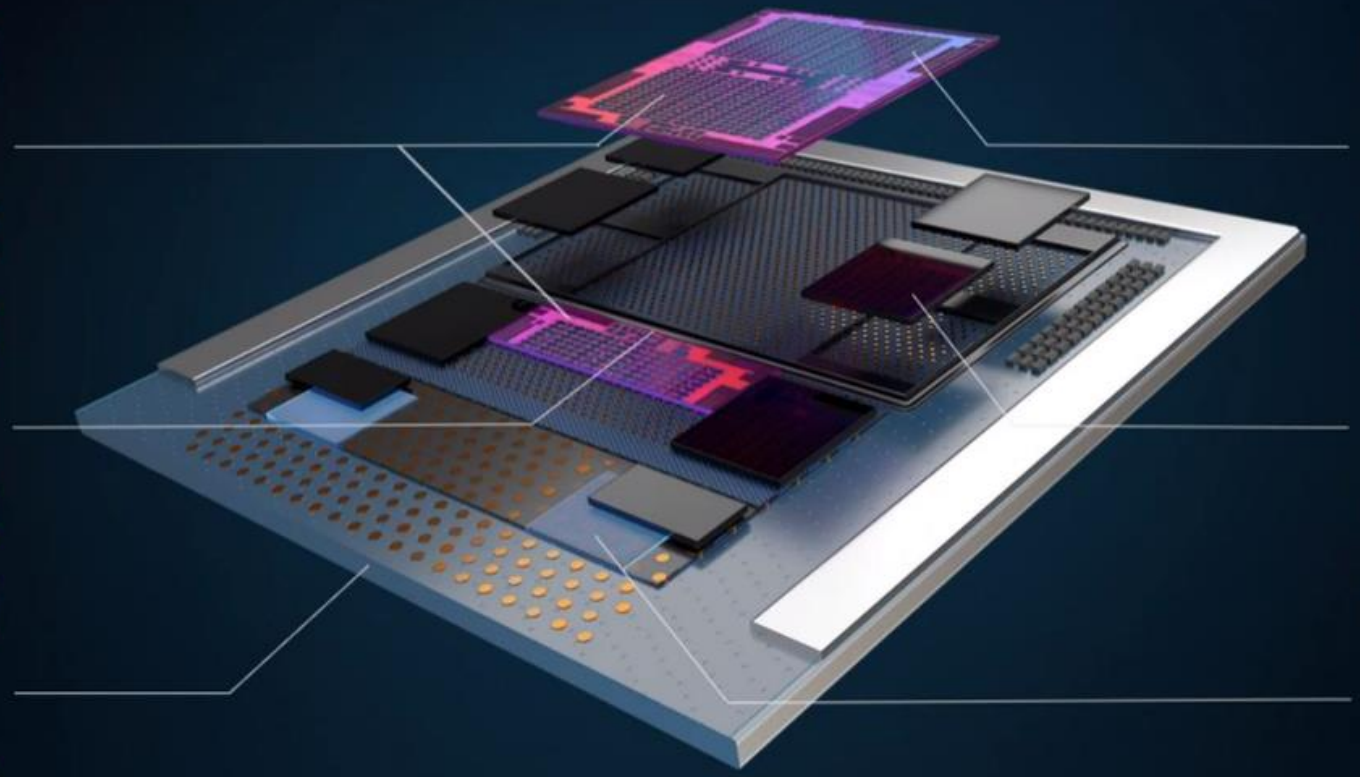
AMD INSTINCT™ MI200 SERIES

KEY INNOVATIONS

TWO
AMD CDNA™2 DIES

ULTRA HIGH BANDWIDTH
DIE INTERCONNECT

COHERENT CPU-TO-GPU
INTERCONNECT



2ND GEN MATRIX
CORES FOR HPC & AI

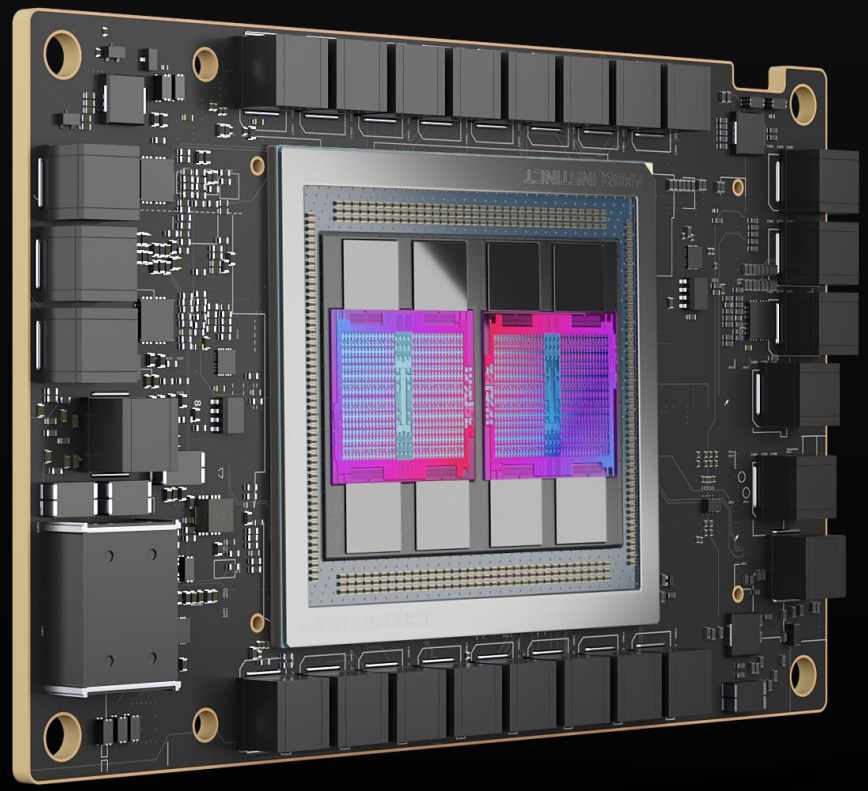
EIGHT STACKS
OF HBM2E

2.5D ELEVATED
FANOUT BRIDGE (EFB)

AMD INSTINCT™ MI200 OAM SERIES

AMD INSTINCT MI250

POWERING DISCOVERIES AT EXASCALE



- 2 AMD CDNA™ 2 GRAPHICS COMPUTE DIES
- 208 COMPUTE UNITS
- 832 MATRIX CORES
- 128GB HBM2E 3.2TB/s BANDWIDTH
- UP TO 6 GPU-TO-GPU INFINITY FABRIC LINKS
- 2 PCIE GPU-TO-GPU LINKS

AMD

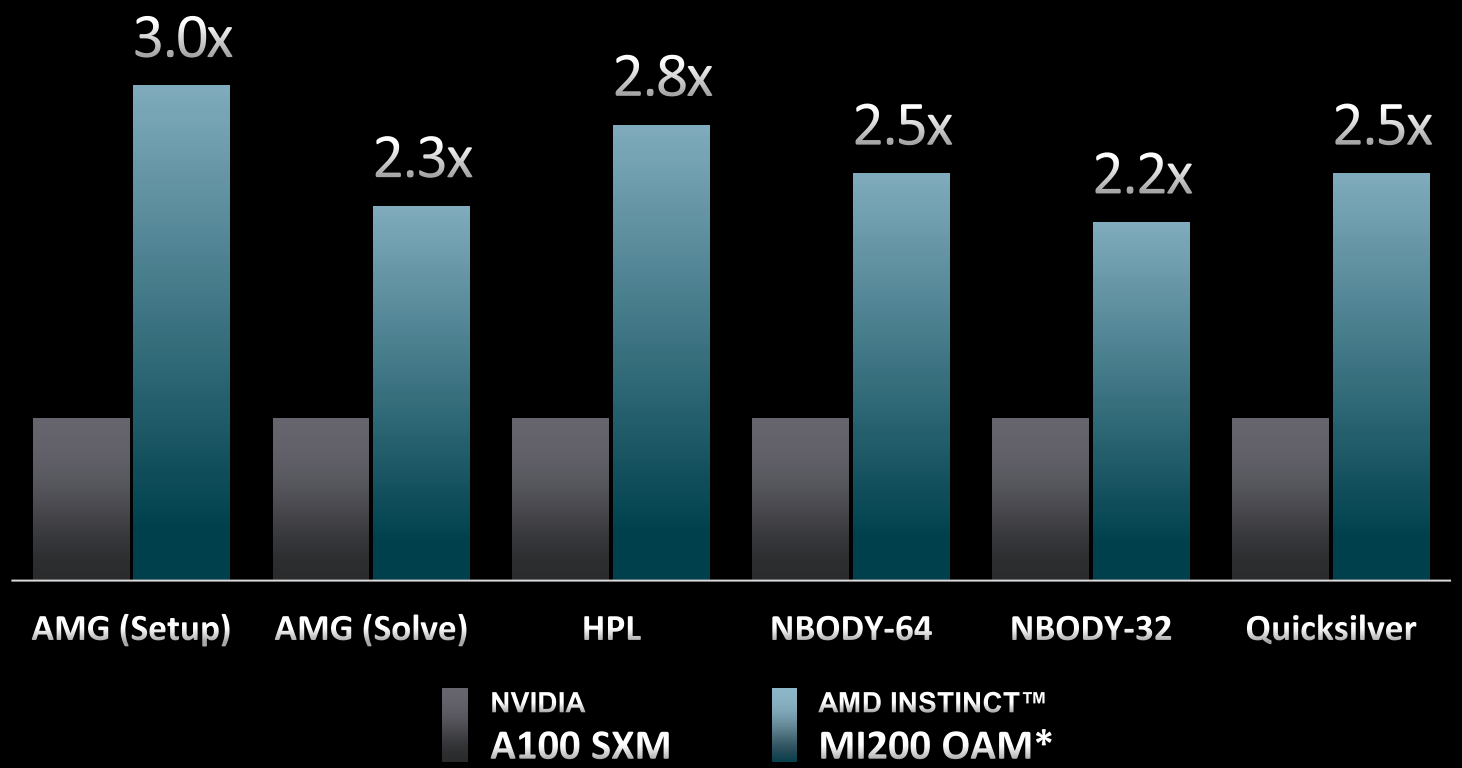
SHATTERING PERFORMANCE BARRIERS IN HPC & AI

PEAK PERFORMANCE	A100	MI200*	INSTINCT™ ADVANTAGE
FP64 VECTOR	9.7 TF	47.9 TF	4.9X
FP32 VECTOR	19.5 TF	47.9 TF	2.5X
FP64 MATRIX	19.5 TF	95.7 TF	4.9X
FP32 MATRIX	N/A	95.7 TF	N/A
FP16, BF16 MATRIX	312 TF	383 TF	1.2X
MEMORY SIZE	80 GB	128 GB	1.6X
MEMORY BANDWIDTH	2.0 TB/s	3.2 TB/s	1.6X

DELIVERING PERFORMANCE FOR HPC

FASTEST HPC APPLICATION PERFORMANCE ACROSS A RANGE OF DOMAINS

HPC BENCHMARKS



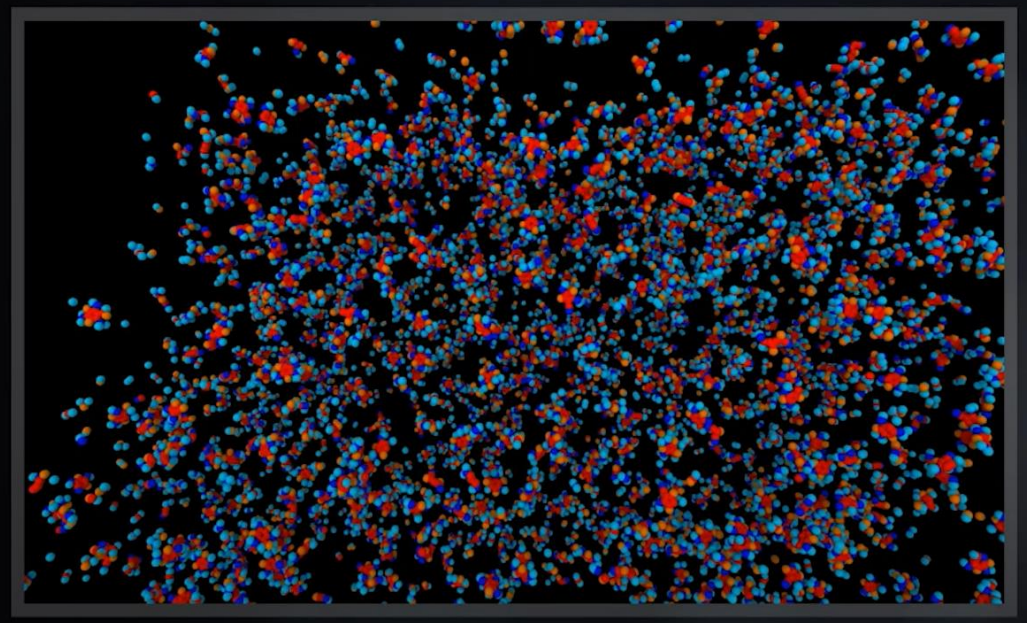
HPC APPLICATIONS

APPLICATION	MI200 ADVANTAGE OVER A100
OpenMM	2.4X
LAMMPS	2.2X
HACC	1.9X
LSMS	1.6X
MILC	1.4X

*MI250X. RESULTS MAY VARY. SEE ENDNOTES: MI200-15A, MI200-16A, MI200-17A, MI200-19A, MI200-20A, MI200-21A, MI200-22A, MI200-23A, MI200-24A, MI200-25A, MI200-26A

LAMMPS COMBUSTION SIMULATION

4 x NVIDIA® A100 SXM



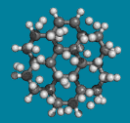
45% COMPLETE

4 x AMD INSTINCT™ MI200 OAM

100% COMPLETE

Top HPC Applications Supported

AMD INSTINCT AMD ROCm



Molecular Dynamics (Available Now)

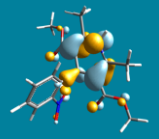
NAMD

LAMMPS

GROMACS

AMBER

OpenMM



Quantum Chemistry (Available soon)

CP2K (Now)

Quantum Espresso

NWChem

VASP

Oil & Gas (Available Now)

Reverse Time Migration

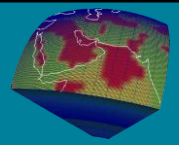
AMD has RTM sample code and in-house experts to help optimize customer codes

Quantum Physics (Available now)

Chroma

MILC

GRID



Other HPC Domains (Available Now)

OpenFoam (Fluid Dynamics)

SPECFEM3D (Geophysics)

Relion (Life Sciences)

ION (Weather)

Machine Learning (Available Now)

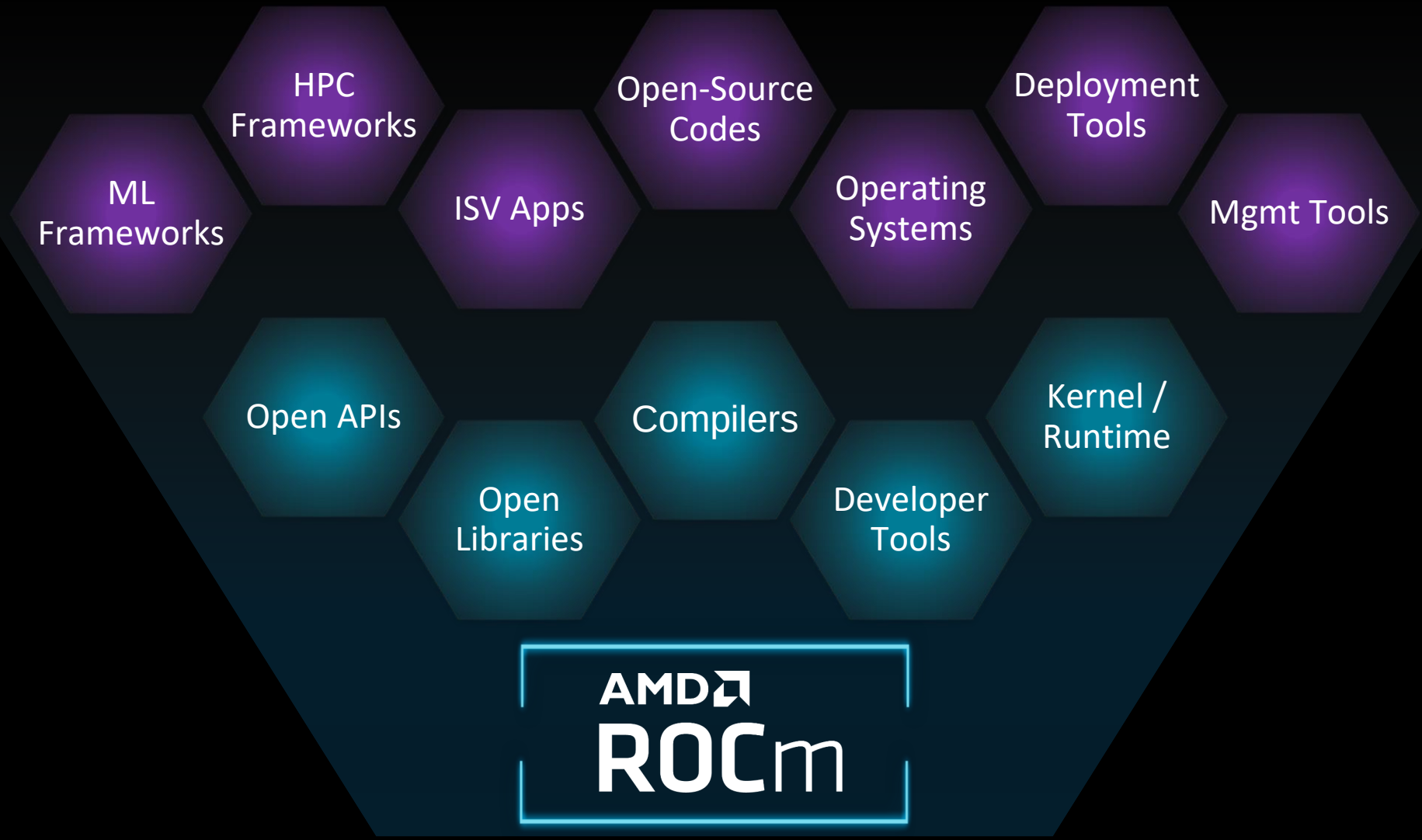
TensorFlow

PyTorch

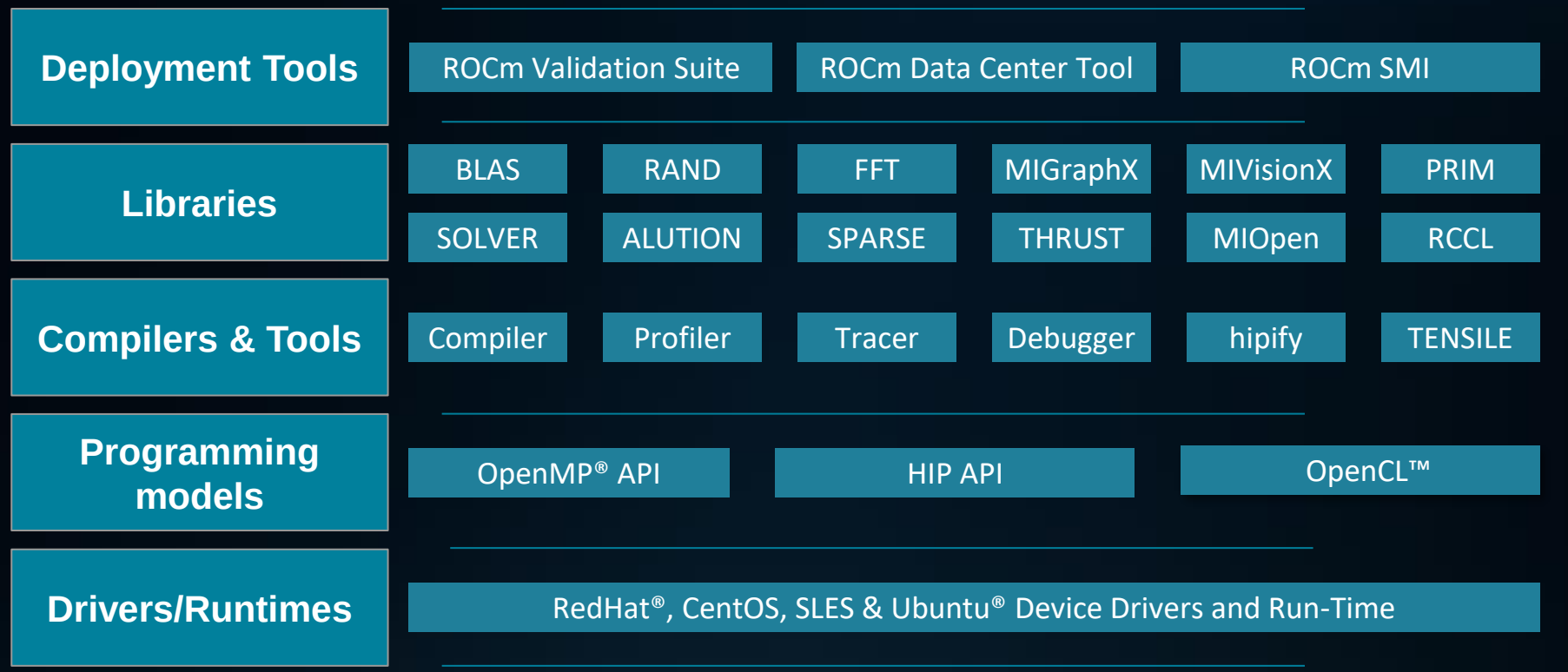
ONNX-Runtime

AMD ROCm – Open Source Software Development Environment

ROCm™: Enabling An Ecosystem Without Borders



AMD ROCm™ - The Core Components



CUDA® Comparable Math & Comm Libraries

CUDA Library	ROCm Library	Description
cuBLAS	rocBLAS	Basic Linear Algebra Subroutines
cuFFT	rocFFT	Fast Fourier Transfer Library
cuSPARSE	rocSPARSE	Sparse BLAS + SPMV
cuSolver	rocSolver	Lapack Library
AMG-X	rocALUTION	Sparse iterative solvers & preconditioners with Geometric & Algebraic MultiGrid
Thrust	rocThrust	C++ parallel algorithms library
CUB	rocPRIM	Low Level Optimized Parallel Primitives
cuDNN	MIOpen	Deep learning Solver Library
cuRAND	rocRAND	Random Number Generator Library
EIGEN	EIGEN	C++ template library for linear algebra: matrices, vectors, numerical solvers
NCCL	RCCL	Communications Primitives Library based on the MPI equivalents

Open Programming Models Supported By ROCm™ Software Platform

HIP

HIP (Heterogeneous Interface for Portability) is an interface that provides similar functionality to CUDA API

- A CUDA-like API that is open-source and portable
- Runtimes for targeting computing on GPU or CPU available
- Developers write once for both GPU or CPU
- GPU runtime included in ROCm™; CPU runtime available on GitHub

OpenMP

OpenMP 5.0, an interface that supports both CPU or GPU shared-memory multiprocessing programming

- Compiles C/C++ code with OpenMP “target” pragmas
- Open-source compiler implementation
- Links with libomptarget to produce a binary that can offload work to the GPU
- GPU runtime included in ROCm; CPU compiler & runtime available separately

CODE CONVERSION TOOLS

EXTEND YOUR APPLICATION PLATFORM SUPPORT BY CONVERTING CUDA® CODE

SINGLE SOURCE

MAINTAIN PORTABILITY

MAINTAIN PERFORMANCE

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

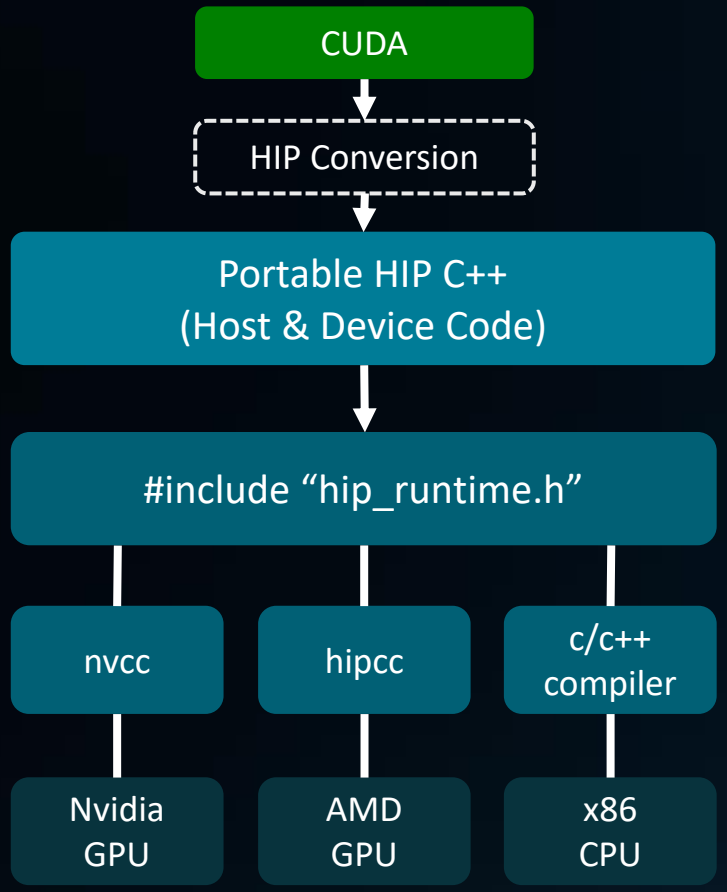
Hipify-clang

- 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+

HIP: High-Performance, Open, and Portable



“Altair AcuSolve CFD, built on a modern framework for GPU computing, was *ported in less than a month* to the open-source AMD ROCm platform using an AMD Radeon Pro VII and easily migrated to a server running the new AMD Instinct™ *MI100, delivering a big boost in performance*.”

Yi Chen
Senior Development Manager, Altair AcuSolve

ALTAIR

“On the Nvidia systems, the performance of the HIP and CUDA kernels are *nearly identical*, indicating there are *no performance losses* from the ‘hipification’ process.”

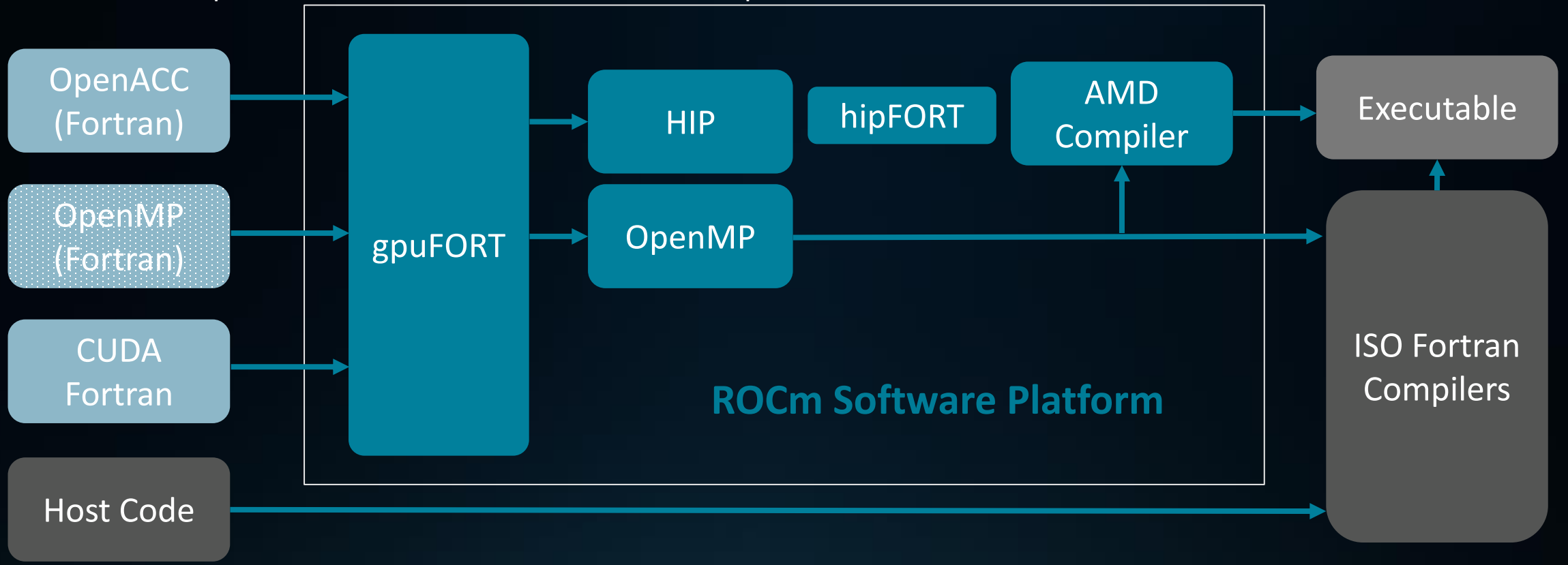
HIP Performance Comparisons: AMD and Nvidia GPUs
https://journal.fluidnumerics.com/hip-performance-comparisons-amd-and-nvidia-gpus

FluidNumerics

GPUFORT is a Source-to-Source Conversion Tool (for Existing FORTRAN Codes)

Conversion tool to translate directive-based code to direct kernel programing source code:

- Fortran+OpenACC and CUDA Fortran -> Fortran + [GCC/AOMP OpenACC/MP runtime calls] + HIP C++
- Fortran+OpenACC and CUDA Fortran -> Fortran + OpenMP 4.5+



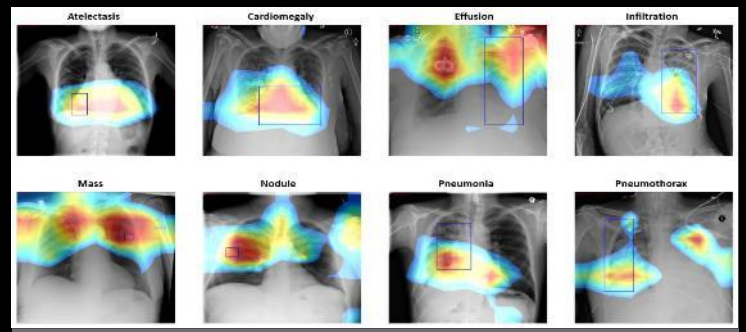
ML FRAMEWORKS & LIBRARIES

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 Q4-2021	Docker Hub	Est 2022
CuPy	cupy.dev	Docker Hub	cupy.dev

Focused on Targeted ML Use-Cases

Most common models on HuggingFace supported on AMD platform today



VIDEO & IMAGE RECOGNITION

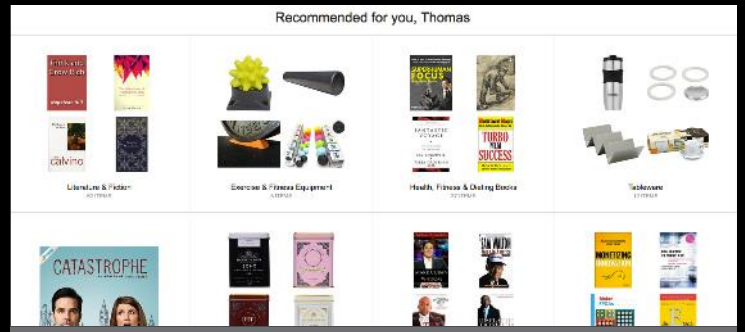
Optimized Models
Resnet, VGG, Inception
GoogleNet, ResNext, Detectron2,
RetinaNet, Mask R-CNN

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

LANGUAGE PROCESSING

Optimized Models
GNMT, BERT, GPT-2,
BART, DeBERTa,
DistilBERT, RoBERTa, T5

Markets
Customer Service
Web Services/E-Commerce



RECOMMENDATION ENGINE

Optimized Models
DLRM

Markets
Web Services/E-commerce
SaaS

ROCm™ Accelerating Science & Discovery

OPEN & PORTABLE

CoMet

"One of the benefits of converting to HIP is that unlike previous CUDA versions of code that could only run on Nvidia GPUs, **the same source code is now portable between GPUs**.

Daniel Jacobson, Computational Systems Biologist, ORNL

LINK

FASTER TIME TO DISCOVERY

PIConGPU

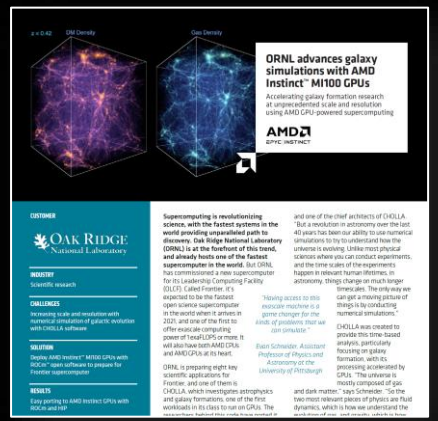
"At every step of the way, we were able to say that **there is a significant increase in performance**. A simulation that took two months on the previous Summit system might take less than two weeks on Frontier…"

Sunita Chandrasekaran, Asst Professor of Computer & Information Sciences at the University of Delaware

LINK

PERFORMANCE AT SCALE

Cholla



"my simulation code runs twice as fast. That means that I can run higher resolution simulations. **Having access to this exascale machine is a game changer** for the kinds of problems that we can simulate.."

Evan Schneider, Asst Professor of Physics and Astronomy at the University of Pittsburgh

LINK

AMD ROCm 5.0

DEMOCRATIZING EXASCALE FOR ALL

EXPANDING SUPPORT & ACCESS

- Support for Radeon™ Pro W6800 Workstation GPUs
- Remote access through the AMD Accelerator Cloud

OPTIMIZING PERFORMANCE

- MI200 Optimizations: FP64 Matrix ops, Improved Cache
- Improved launch latency and kernel performance

ENABLING DEVELOPER SUCCESS

- HPC Apps & ML Frameworks on AMD Infinity Hub
- Streamlined and improved tools to increase productivity

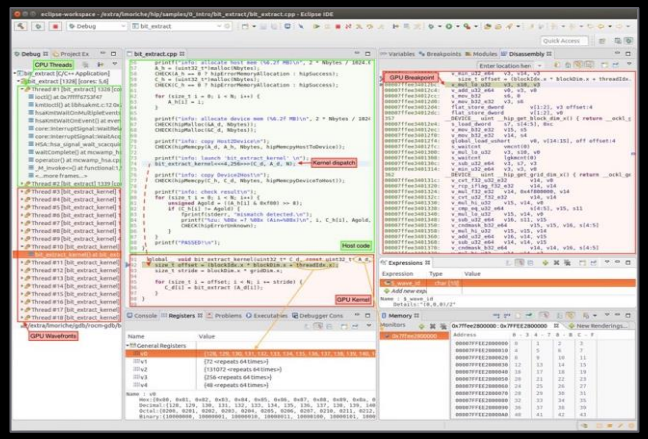
ROCm 5.0: Enabling Developer Success with an Open Toolchain

COMPILER

LLVM based

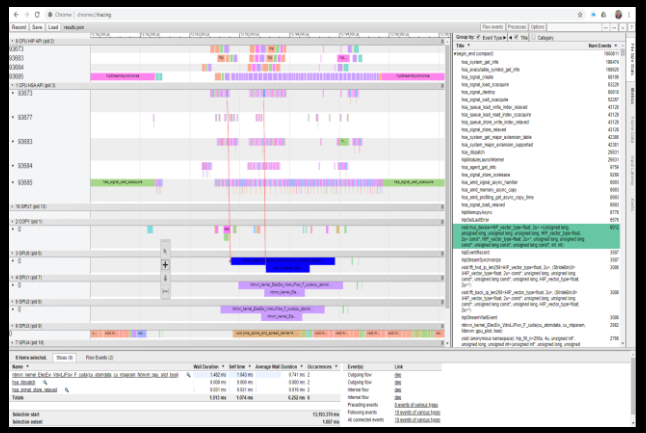
- C/C+ and Fortran languages
- HIP and OpenMP offload
- Compile CPU & GPU code with one tool

DEBUGGER



- rocGDB for CPU and GPU
- Debug source level code for HIP and OpenMP
- GDB-enabled GUI integration

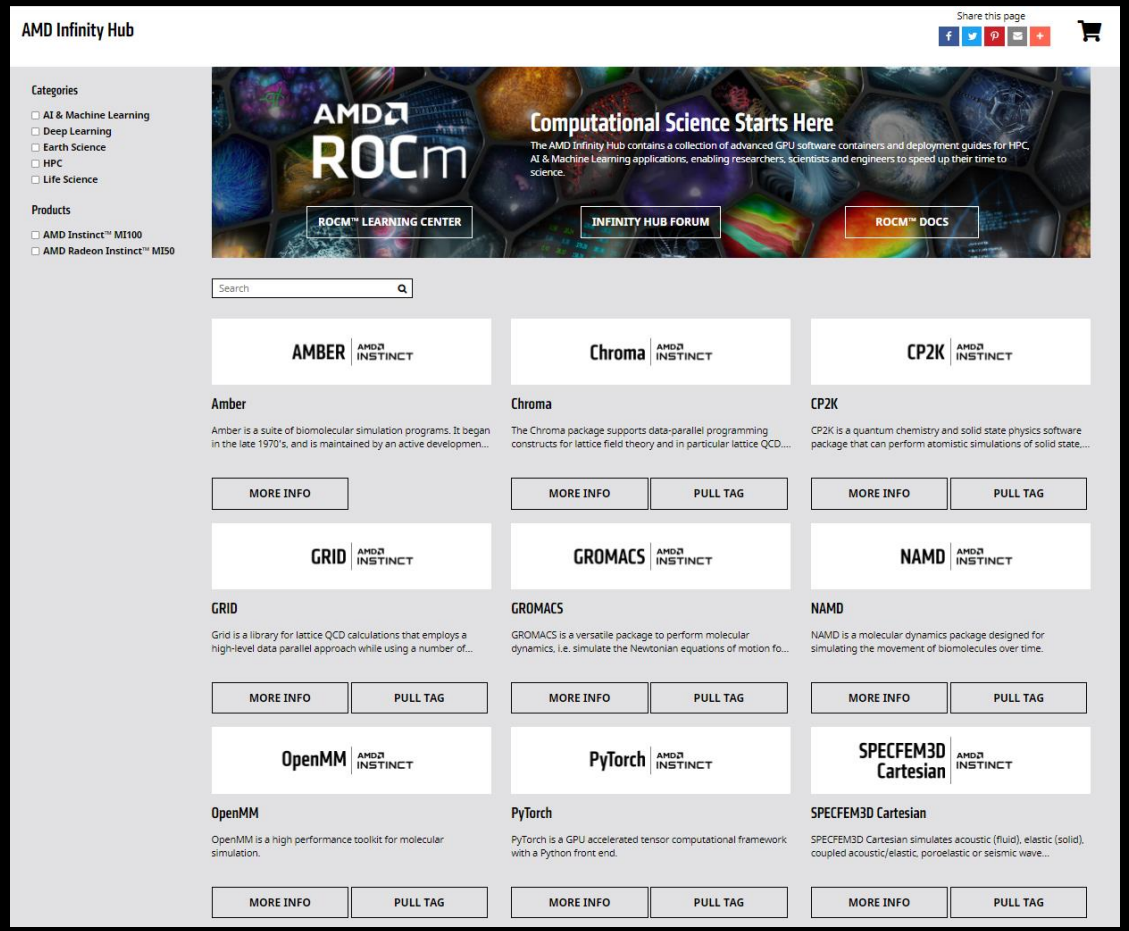
PROFILER



- rocProf API and CLI with JSON output
- 3rd party integrations underway (HPCToolKit, TAU, Vampir, Score-P, ARM Forge, Likwid)

WHAT'S NEW WITH INFINITYHUB

More Apps, More Numbers



MI200 Support

- HPC Apps: CHROMA, CP2K, GRID, GROMACS, HACC, LAMMPS, MILC, NAMD 3.0, OpenMM, Relion, SPECFEM3D (Cartesian), SPECFEM3D (Globe)
- Benchmarks: HPL, NBODY

Additional MI200 Support Planned for 1H22

- HPC Apps: AMBER*, ICON, MPAS, NWCHEM, OpenFOAM, PYFR, QuantumEspresso, WRF, NEMO
- Benchmarks: MLPerf (SSD, Resnet50, Transformer), HPCG

Performance Results for Select Apps / Benchmarks

https://www.amd.com/en/graphics/server-accelerators-benchmarks

** Available on InfinityHub with MI100 support today*

Getting Started with ROCm™ Open Software Platform

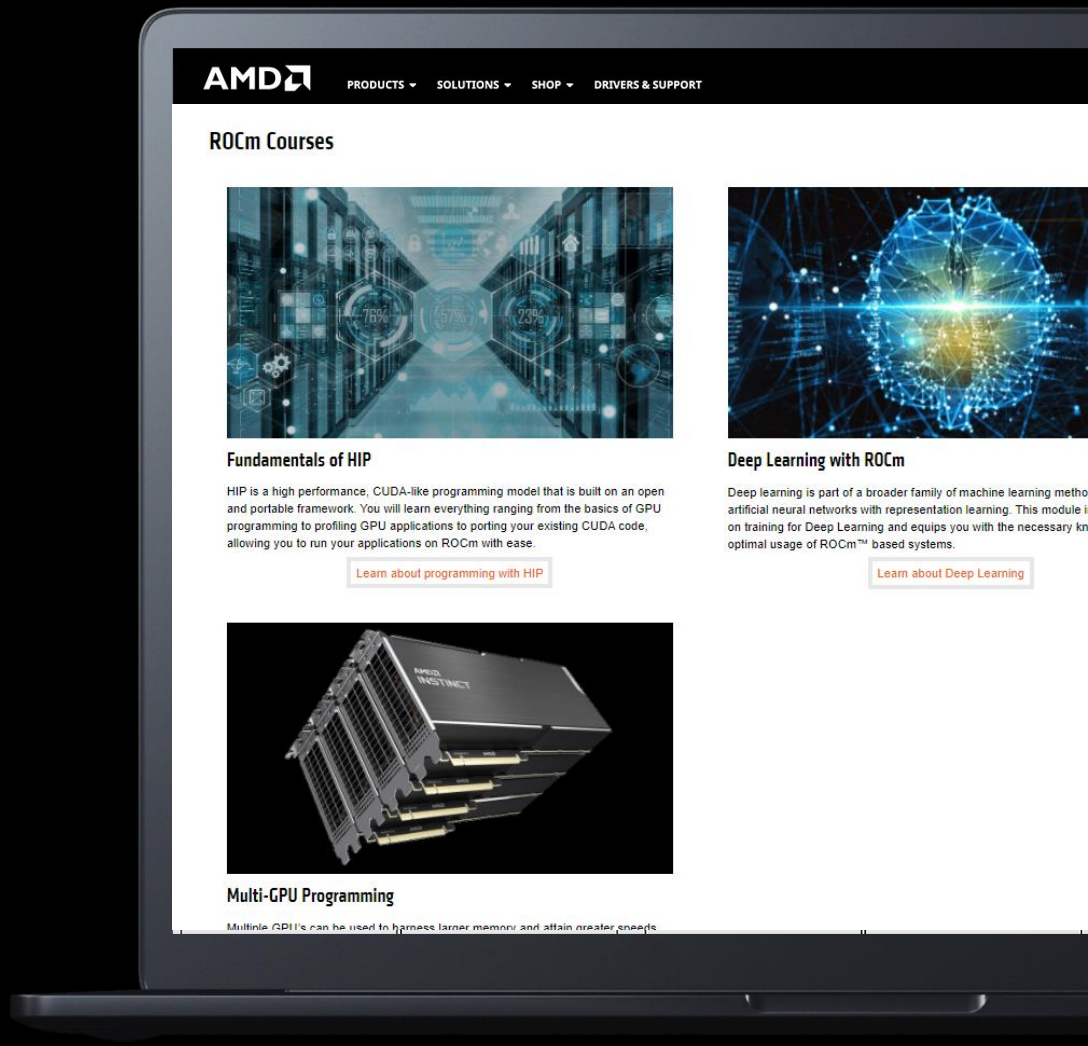
ROCm™ Learning Center

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



CONVERGENCE OF HPC & AI

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

JAX | DEEPSPEED | CUPY

AMD GPU and ROCm Resources

AMD Instinct MI200 GPU and ROCm References

MI200 Product Video:
https://youtu.be/Bm2r4Z7qlcs

MI200 Brochure:
https://www.amd.com/system/files/documents/amd-instinct-mi200-datasheet.pdf

CDNA 2 Whitepaper:
https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf

LAMMPS Combustion Simulation Video: https://youtu.be/zEKooI1UXYA

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

AMD MI250x OAM for Exascale: https://www.amd.com/en/products/server-accelerators/instinct-mi250

AMD MI250 OAM: AMD Instinct™ MI250 Accelerator | AMD

AMD HPC and AI Resources

Software & Documents

AMD Infinity Hub: ROCm™ containers, learning center (AMD.com/InfinityHub)

AMD Developer Hub: Spack recipes, compiler, math libraries, µProf, and additional resources (developer.amd.com)

AMD Tech Docs and Whitepapers

HPC Tuning Guide for AMD EPYC™ 7003 Series Processor

AMD Initiatives

AMD Covid-19 HPC Fund

AMD Energy Efficiency goals for HPC and AI

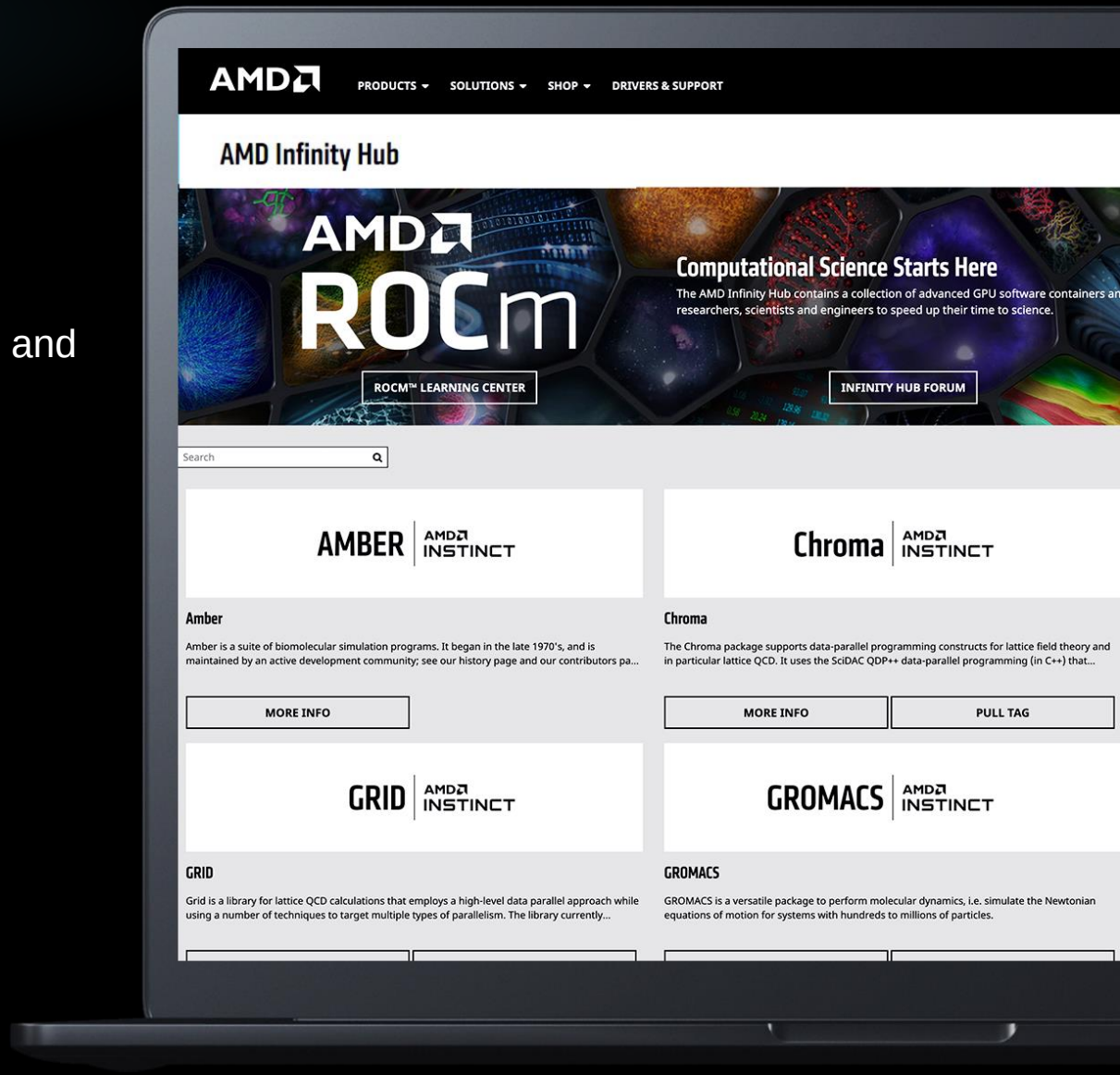
AMD Instinct™ Education & Research Initiative

AMD HPC Users Forum

End user community forum driven by users for users

AMD Instinct Benchmarks

AMD Instinct™ Benchmarks | AMD




AMD