



# ***AMD INSTINCT GPU For AI and HPC***

***DC GPU BU: 沈 仲 杰***



# AMD PLATFORM FOR ACCELERATED COMPUTING

LEADERSHIP IN HPC & AI FOR EXASCALE-CLASS COMPUTING



**WORKLOAD-OPTIMIZED  
COMPUTE ARCHITECTURE**

PURPOSE-BUILT, OPTIMIZED  
ARCHITECTURE  
DESIGNED TO DO ONE THING  
EXTREMELY WELL: COMPUTE  
INTENSIVE HPC AND AI  
WORKLOADS

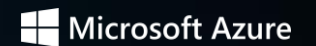
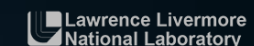


**OPEN & PORTABLE  
SOFTWARE**

SIMPLIFIED PROGRAMMABILITY  
AND INCREASED USABILITY  
THROUGH A GROWING SOFTWARE  
ECOSYSTEM AND A RICH SUITE OF  
OPTIMIZED LIBRARIES,  
FRAMEWORKS AND TOOLS



**DEEP PARTNERSHIP WITH  
LEADING HPC CENTERS & AI  
THOUGHT-LEADERS**





# LEADING THE EXASCALE ERA

- Powering World's #1 Supercomputer  
First to break Exascale barrier
- Powering 75% of World's Top 20 Green Supercomputers  
15 of top 20 most efficient systems rely on AMD
- Powering World's #1 HPL-MxP Supercomputer  
6.86 Eflops on HPL-MxP Mixed-Precision Benchmark
- 38% growth in TOP500 systems Year-over-Year  
Powering 101 of Top500 systems



# SHATTERING PERFORMANCE BARRIERS IN HPC & AI



## PEAK PERFORMANCE

*A100*

*MI250*

*INSTINCT*  
ADVANTAGE

FP64 VECTOR	9.7 TF	45.3 TF	<b>4.6X</b>
FP32 VECTOR	19.5 TF	45.3 TF	<b>2.3X</b>
FP64 MATRIX	19.5 TF	90.5 TF	<b>4.6X</b>
FP32 MATRIX	N/A	90.5 TF	N/A
FP16 MATRIX	312 TF	362 TF	<b>1.2X</b>
MEMORY SIZE	80 GB	128 GB	<b>1.6X</b>
PEAK MEMORY BANDWIDTH	2.0 TB/s	3.2 TB/s	<b>1.6X</b>

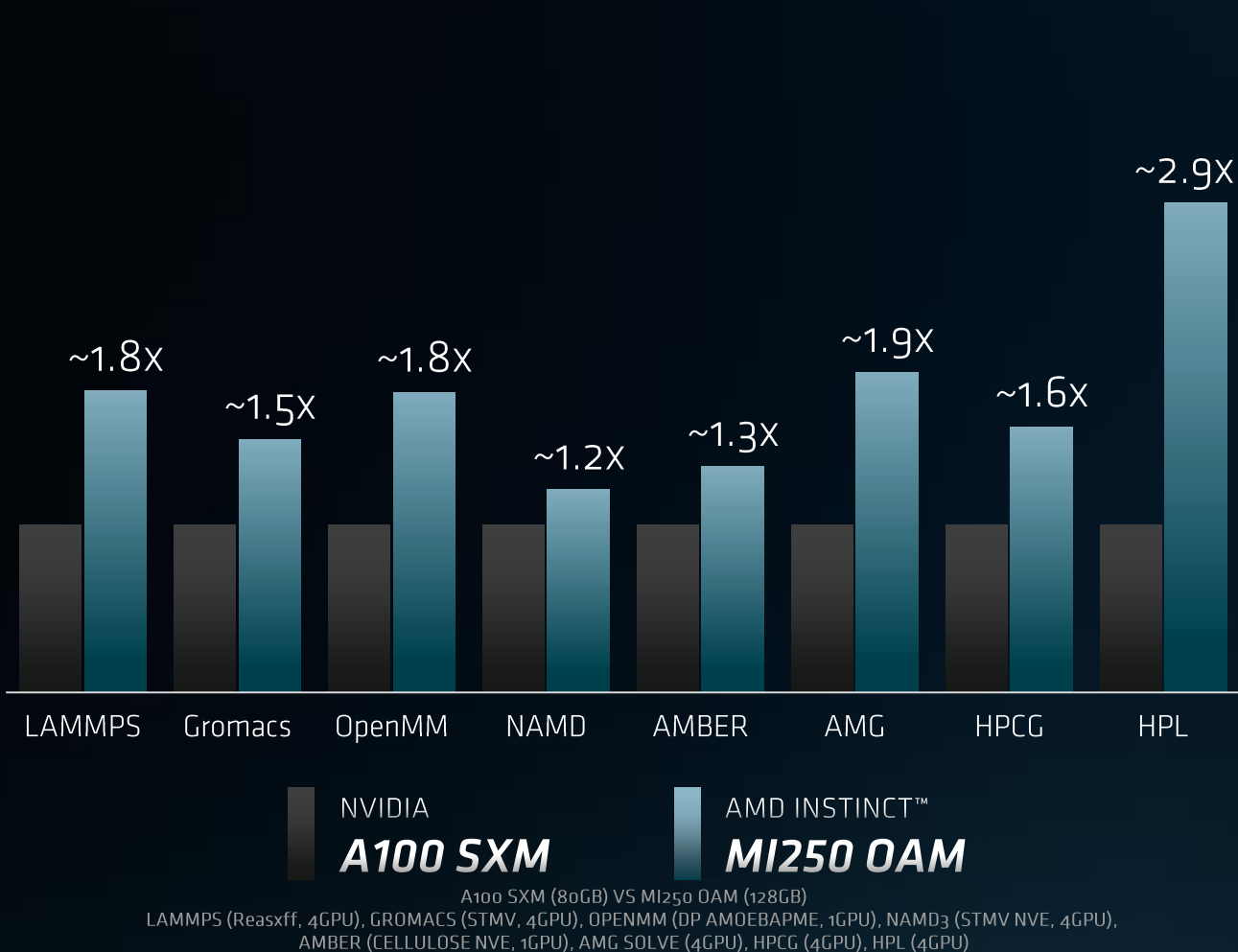
NOTE: THE A100 TF32 DATA FORMAT IS NOT IEEE FP32 COMPLIANT, SO NOT INCLUDED IN THIS COMPARISON.

SEE ENDNOTES: MI200-01, MI200-07



# MI250 – BEST PERFORMANCE IN HPC

## FASTEST HPC PERFORMANCE ACROSS A RANGE OF DOMAINS



*GROMACS OpenMM LAMMPS NAMD AMBER Relion*

*Hoomd-Blue MILC NWChem OpenFOAM PETSc*

2022 *PyTorch TensorFlow ONNX JAX*

*Ansys Mechanical Cascade Charles Altair AcuSolve*

*Tempoquest WRF/AceCast + MORE*

*VTKm HACC AMG Solve BDAS Quantum Espresso*

*VASP QMCPack Quicksilver Nekbone PENNANT*

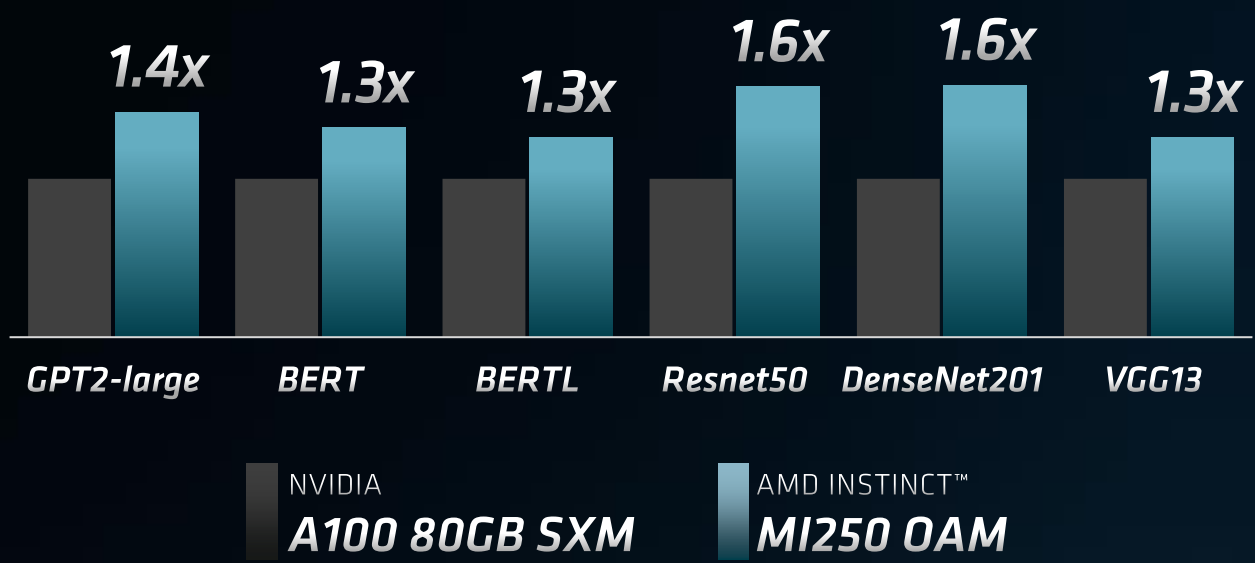
2023 *Kripke*

*Ansys Fluent Dassault XFlow Dassault CST STAR-CCM+*

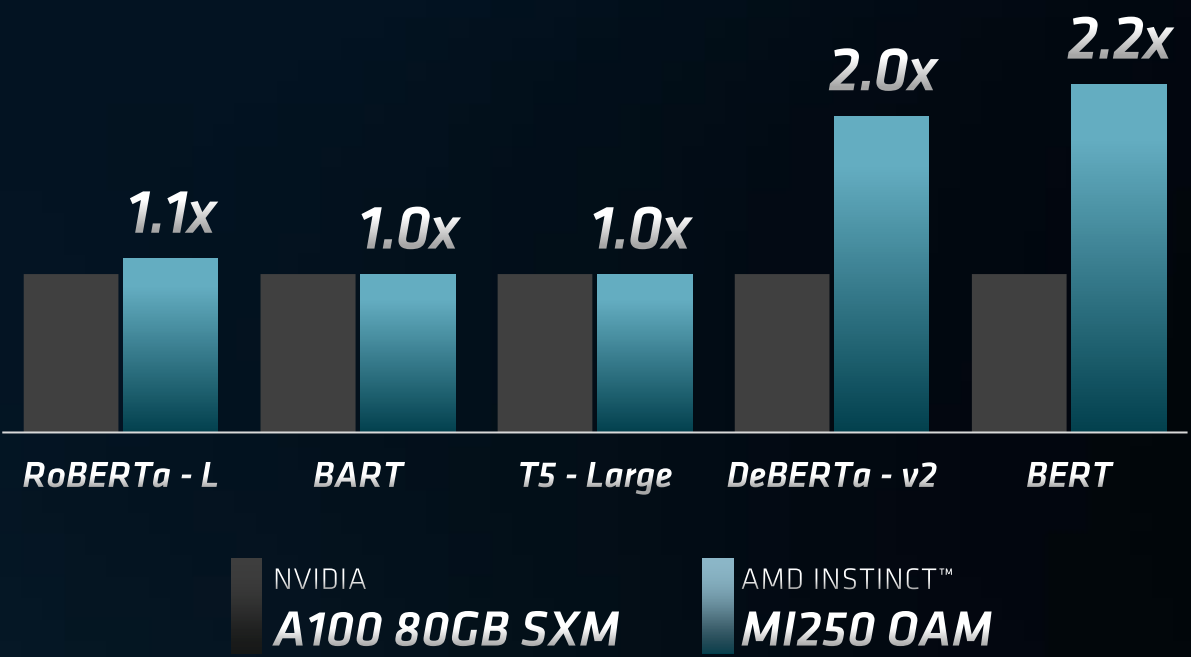
*Devito Pro + MORE*

# ML TRAINING – PARTNERING WITH INDUSTRY LEADERS

MICROSOFT SUPERBENCH



HUGGING FACE



A100 SXM (80GB) VS MI250 OAM (128GB)  
PERF #'S AS OF 11/09/22

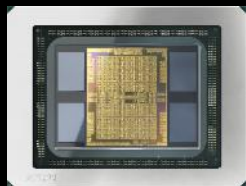


JAX

DEEPSPEED

CUPY

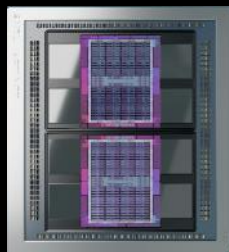
# OUR JOURNEY IN GPU ACCELERATION



AMD Instinct™ **MI100**  
AMD CDNA™

## *Ecosystem Growth*

First purpose-built GPU  
architecture for the data center



AMD Instinct™ **MI200**  
AMD CDNA™<sub>2</sub>

## *Driving HPC and AI to a New Frontier*

First purpose-built GPU powering  
discovery at Exascale



AMD Instinct™ **MI300**  
AMD CDNA™<sub>3</sub>

## *Data Center APU*

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

2020

2024

# AMD INSTINCT™ MI300A

The world's first data center integrated CPU + GPU



Next-Gen  
Accelerator  
Architecture



24 Leadership  
Data Center  
CPU cores

**146B**

Transistors

**128GB**

HBM3

**3D**

Advanced Chiplet Packaging



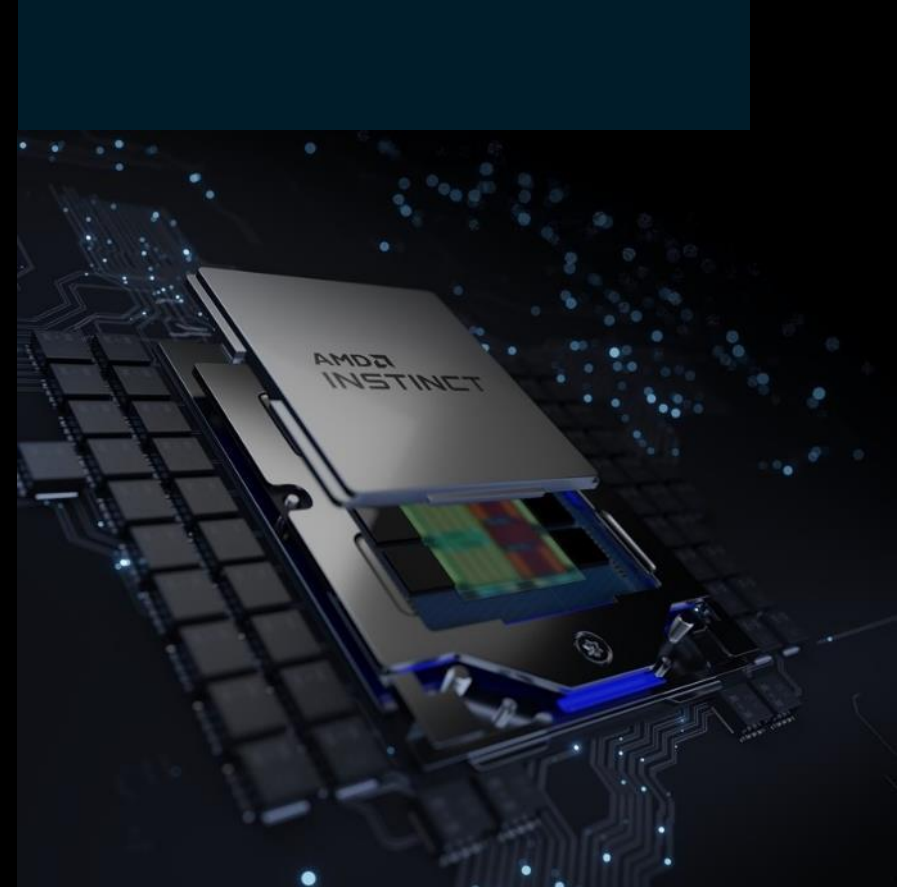
# AMD INSTINCT™ MI300

## THE WORLD'S FIRST DATA CENTER INTEGRATED CPU + GPU

- 4<sup>th</sup> 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

---

***Planned for 2H 2023***



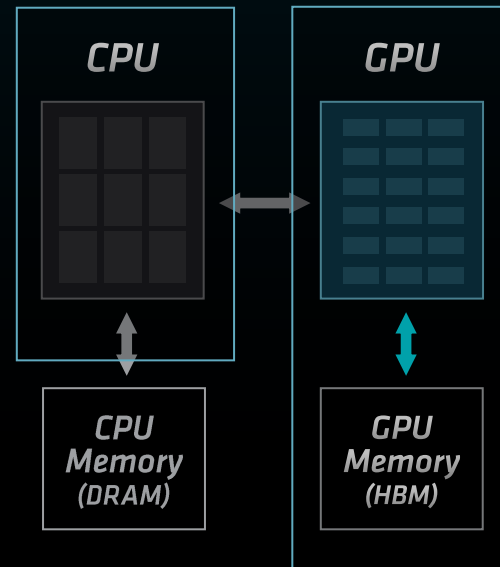
**> 8X**

***Expected AI Training Performance  
vs. MI250X***

# UNIFIED MEMORY APU ARCHITECTURE BENEFITS

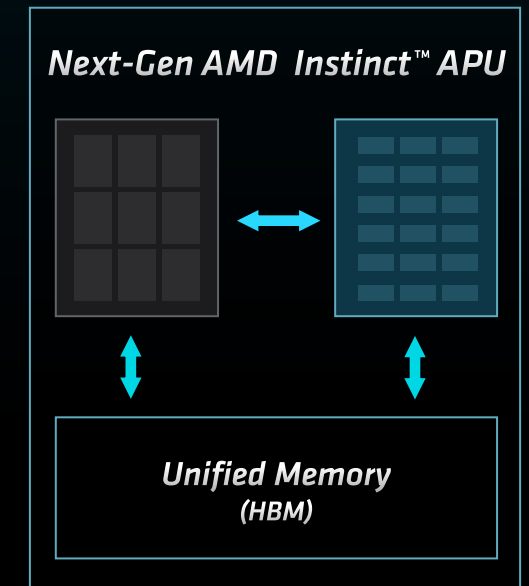
## AMD CDNA™ 2 Coherent Memory Architecture

- 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 4<sup>th</sup> Gen Infinity Interconnect
- Low TCO with Unified Memory APU Package

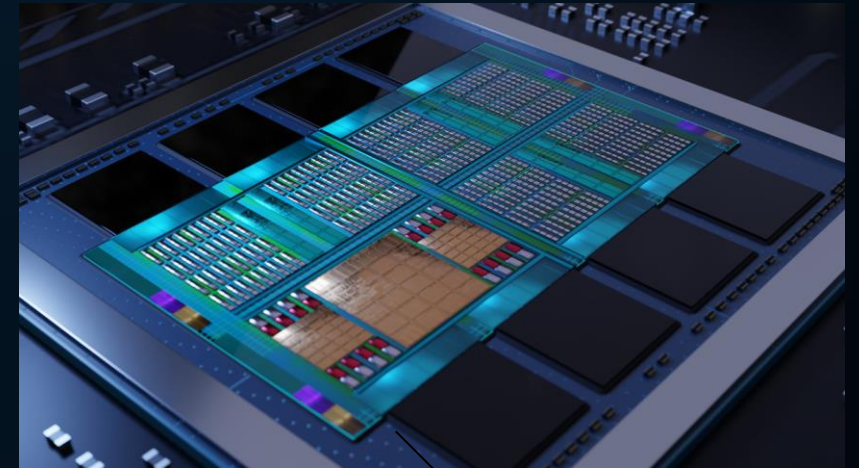


# AMD INSTINCT MI300A ARCHITECTURE: AN INSIDE VIEW

- CPU-GPU HW enabled coherent with hybrid 3D bonding
- APU-APU coherent with Infinity Fabric interconnect



- ✓ HW support for TF32, FP8, and ML sparsity
- ✓ Eliminate expensive data copy operations
- ✓ Reduced kernel launch latency
- ✓ Co-execution of integer & FP operations
- ✓ Transparent management of CPU and GPU caches via cache coherence
- ✓ Task offloads by both CPU cores and GPU CU to each other or other CPU/GPU units
- ✓ Efficient synchronization mechanisms
- ✓ Workload optimized, open, and portable with ROCm support

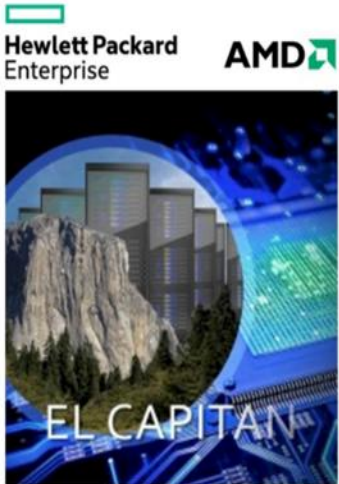




# LEADING THE EXASCALE ERA AGAIN



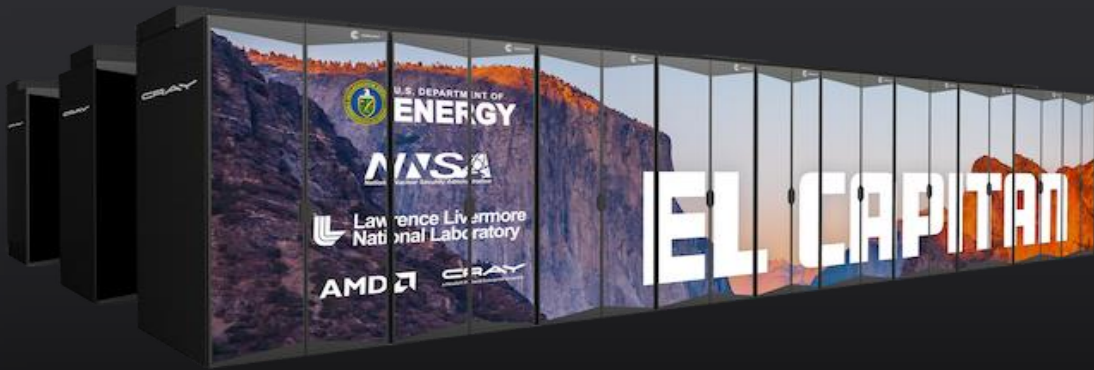
NNSA's El Capitan, is designed to give the U.S. a competitive edge in national security



## El Capitan Features

- Greater than a 10 X increase in performance
- Theoretical Peak  $\geq 2.0$  DP exaflops
- Peak power < 40 MW
- AMD MI -300 APU - 3D chiplet design w/ AMD CDNA3 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

Content provided by LLNL



Expected Late 2023



Introducing today

# *AMD Instinct<sup>™</sup> Platform*

**8x** MI300X

**1.5 TB** HBM<sub>3</sub> Memory

Industry-Standard Design



# AMD 5.x ROCm

DEMOCRATIZING EXASCALE FOR ALL

## EXPANDING SUPPORT & ACCESS

- Widely available in OEM & ODM GPU servers
- Remote access through the AMD Accelerator Cloud

## OPTIMIZING PERFORMANCE

- ML200 Optimizations: FP64 Matrix ops, Improved Cache use
- Improved math and communication libraries

## ENABLING DEVELOPER SUCCESS

- HPC Apps & ML Frameworks on AMD Infinity Hub
- Helps boost productivity with streamlined & improved toolchain



# OPEN SOFTWARE PLATFORM FOR GPU COMPUTE



- Unlocked GPU Power To Accelerate Computational Tasks
- Optimized for HPC and Deep Learning Workloads at Scale
- Open Source Enabling Innovation, Differentiation, and Collaboration

*Benchmarks & App Support*

*Operating Systems Support*

*Cluster Deployment*

*Framework Support*

*Libraries*

*Programming Models*

*Development Toolchain*

*Drivers & Runtime*

*Deployment Tools*

*Optimized Training/Inference Models & Applications*

*MLPERF*

*HPL/HPCG*

*Life Science*

*Geo Science*

*Physics*

*RHEL*

*CentOS*

*SLES*

*Ubuntu®*

*Singularity*

*Kubernetes®*

*Docker®*

*SLURM*

*Kokkos/RAJA*

*PyTorch*

*TensorFlow*

*BLAS*

*RAND*

*FFT*

*MIGraphX*

*MIVisionX*

*PRIM*

*SOLVER*

*ALUTION*

*SPARSE*

*THRUST*

*MIOpen*

*RCCL*

*OpenMP® API*

*OpenCL™*

*HIP API*

*Compiler*

*Profiler*

*Tracer*

*Debugger*

*hipify*

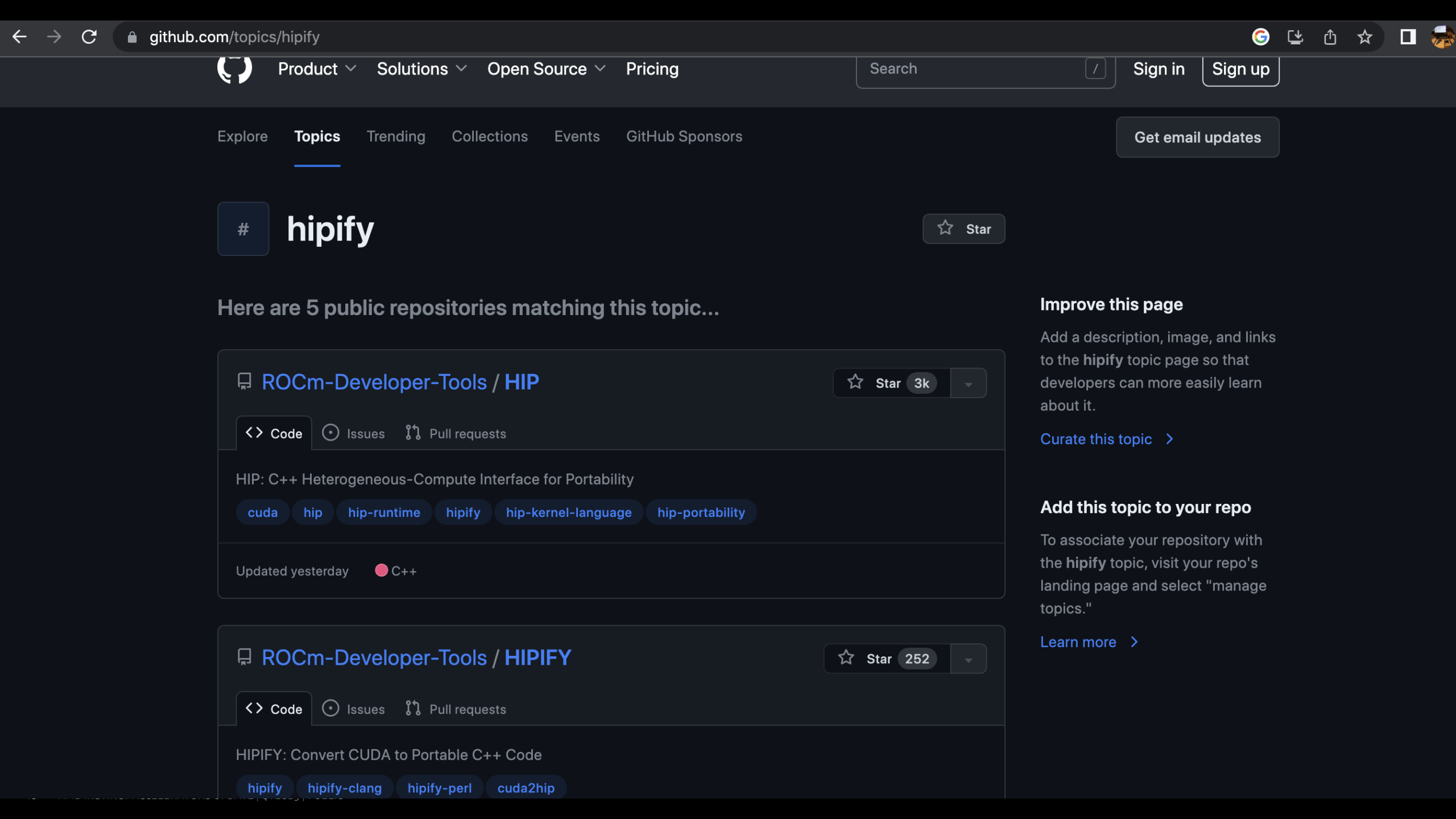
*GPUFort*

*GPU Device Drivers and ROCm Run-Time*

*ROCm Validation Suite*

*ROCm Data Center Tool*

*ROCm SMI*



Product ▾

Solutions ▾

Open Source ▾

Pricing

Search

/

Sign in

Sign up

Explore

Topics

Trending

Collections

Events

GitHub Sponsors

Get email updates

#

hipify



Star

Here are 5 public repositories matching this topic...



ROCm-Developer-Tools / HIP



Star

3k



Code



Issues



Pull requests

HIP: C++ Heterogeneous-Compute Interface for Portability

cuda

hip

hip-runtime

hipify

hip-kernel-language

hip-portability

Updated yesterday



C++



ROCm-Developer-Tools / HIPIFY



Star

252



Code



Issues



Pull requests

HIPIFY: Convert CUDA to Portable C++ Code

hipify

hipify-clang

hipify-perl

cuda2hip

### Improve this page

Add a description, image, and links to the **hipify** topic page so that developers can more easily learn about it.

[Curate this topic](#) >

### Add this topic to your repo

To associate your repository with the **hipify** topic, visit your repo's landing page and select "manage topics."

[Learn more](#) >

# 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
Quicksilver	Benchmark	Now
MPAS	Weather	Now
PETSc	Library	Now
Ansys® Mechanical™ *	ISV	1H' 2023
OpenFOAM®	CFD	1H' 2023
ICON	Weather	1H' 2023
Hoomd-Blue	Life Science	1H' 2023
NWCHEM	Quantum Chemistry	1H' 2023
AceCast (Based on WRF)	ISV	1H' 2023
HPL-AI	Benchmark	1H' 2023
SHOC	Benchmark	1H' 2023
BabelStream	Benchmark	1H' 2023
Cholla	CFD	1H' 2023
Relion v4	Cryo-EM	1H' 2023
Quantum Espresso	Physics	1H' 2023
Trilinos	Solvers, Framework	1H' 2023
VASP	Life Science	2H' 2023

\*Commercial SW with links to ported/optimized code or application with targeted availability – dates subject to change

\*\*Two containers, one for MI100 and one for MI200



## INSTALL PYTORCH




Select your preferences and run the install command. Stable represents the most currently tested and supported version of PyTorch. This should be suitable for many users. Preview is available if you want the latest, not fully tested and supported, builds that are generated nightly. Please ensure that you have **met the prerequisites below (e.g., numpy)**, depending on your package manager. Anaconda is our recommended package manager since it installs all dependencies. You can also [install previous versions of PyTorch](#). Note that LibTorch is only available for C++.

PyTorch Build	Stable (2.0.0)		Preview (Nightly)	
Your OS	Linux	Mac	Windows	
Package	Conda	Pip	LibTorch	Source
Language	Python		C++ / Java	
Compute Platform	CUDA 11.7	CUDA 11.8	ROCm 5.4.2	CPU
Run this Command:	<pre>pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/rocm5.4.2</pre>			

**NOTE:** PyTorch LTS has been deprecated. For more information, see [this blog](#).

# ML FRAMEWORKS & LIBRARIES

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

	Source	Container	PIP Wheel
 <b>TensorFlow</b>	<a href="#">TensorFlow GitHub</a>	<a href="#">Infinity Hub</a>	<a href="#">pypi.org</a>
 <b>PyTorch</b>	<a href="#">PyTorch GitHub</a>	<a href="#">Infinity Hub</a>	<a href="#">pytorch.org</a>
 <b>ONNX RUNTIME</b>	<a href="#">ONNX-RT GitHub</a>	<a href="#">Docker Instructions</a>	<a href="#">onnxruntime.ai</a>
<b>JAX</b>	<a href="#">GitHub public fork</a>	<a href="#">Docker Hub</a>	<i>Est Q2 2023</i>
<b>DeepSpeed</b>	<a href="#">DeepSpeed GitHub</a>	<a href="#">Docker Hub</a>	<a href="#">deepspeed.ai</a>
<b>CuPy</b>	<a href="#">cupy.dev</a>	<a href="#">Docker Hub</a>	<a href="#">cupy.dev</a>

PYTORCH FOUNDATION

FOUNDING MEMBER

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.  
PyTorch, the PyTorch logo and any related marks are trademarks of Facebook, Inc.



# rocm/deepspeed ☆

By [rocm](#) • Updated 10 months ago

Image

↓ Pulls 792

OverviewTags

 Build passing pypi package 0.9.4 docs passing License MIT downloads/month 260k

DeepSpeed+Megatron trained the world's most powerful language model: [MT-530B](#)

DeepSpeed is hiring, [come join us!](#)


DeepSpeed is a deep learning optimization library that makes distributed training easy, efficient, and effective.

10x Larger Models

10x Faster Training

## Docker Pull Command

```
docker pull rocm/deepspeed
```



## Source Repository

 Github



ROCm Documentation is transitioning to this site. For the legacy documentation, please visit [docs.amd.com](#). For more information or to provide feedback about this documentation transition, please see [our announcement](#).

ROCm Documentation

Home

What is ROCm?

Deploy ROCm

Linux Quick Start

Linux Overview

▼

Docker

Release Info

Release Notes

Changelog

GPU Support and OS Compatibility (Linux)

Known Issues [↗](#)

Compatibility

▼



# AMD ROCm™ Documentation

Applies to Linux    📅 2023-05-25    ⌚ 4 min read time

- What is ROCm? ▼
- Deploy ROCm ▼
- Release Info ▼

## APIs and Reference

- [Compilers and Development Tools](#)
- [HIP](#)
- [OpenMP](#)
- [Math Libraries](#)
- [C++ Primitives Libraries](#)
- [Communication Libraries](#)

## Understand ROCm

- [Compiler Disambiguation](#)
- [Using CMake](#)
- [Linux Folder Structure Reorganization](#)
- [GPU Isolation Techniques](#)
- [GPU Architecture](#)

# AMD ROCm™

## LEARNING CENTER

GET UP TO SPEED ON ROCm™

### GETTING STARTED

Understand the key components of ROCm™ and be set up for success when you start developing

### ON-DEMAND ROCm™ COURSES & LIVE TRAINING SESSIONS

Learn the fundamentals of HIP, multi-GPU Programming, and Deep Learning with ROCm™ through both on-demand video and monthly live training

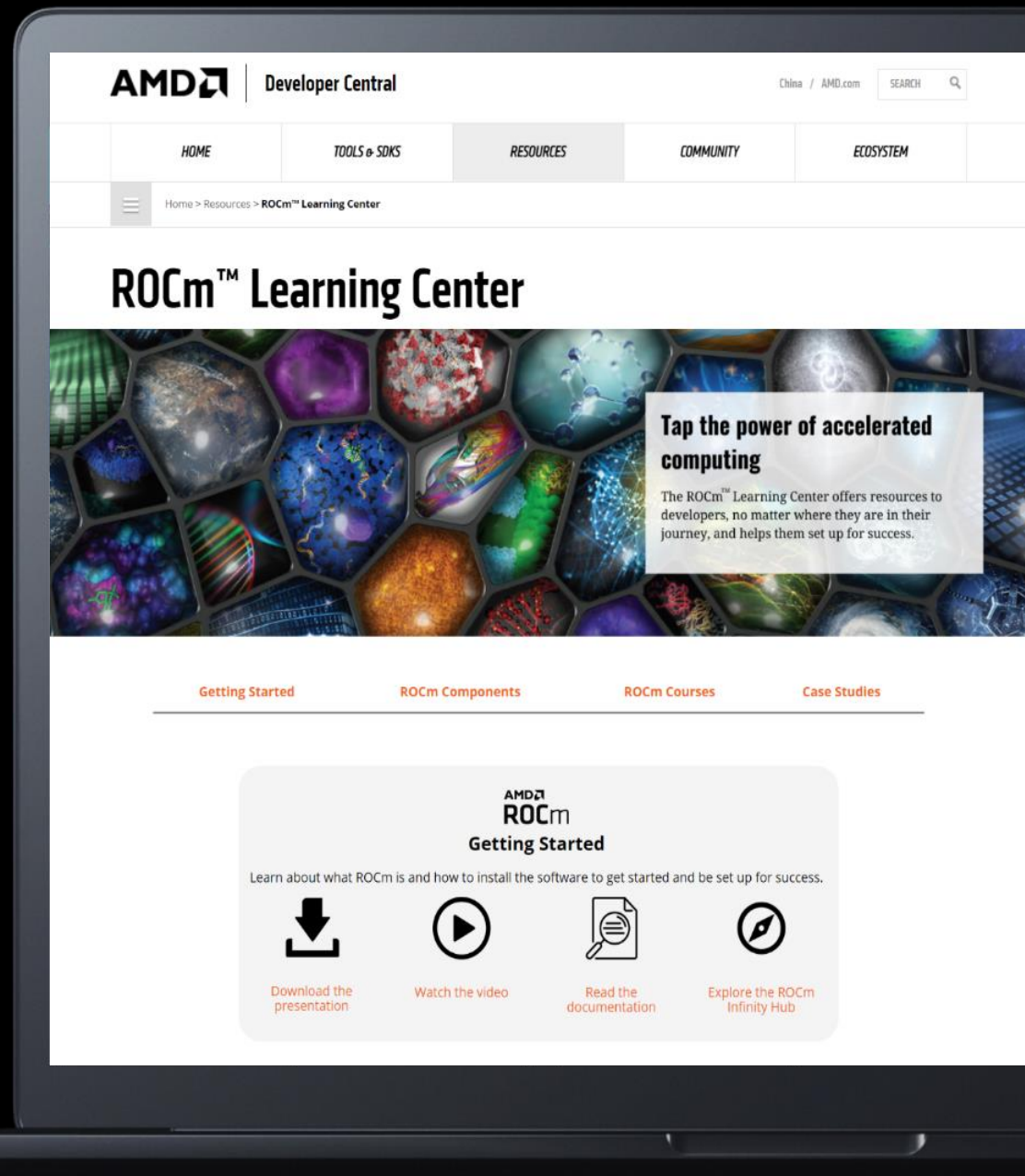
### USER COMMUNITY

See real world applications through case studies and learn through hands-on trainings with our lunch and learn series

### TAP THE POWER OF ACCELERATED COMPUTING

The ROCm Learning Center offers resources to developers, no matter where they are in their journey, and helps set them up for success.

[LEARN MORE](#)



# AMD INFINITY HUB

*Turnkey solutions for HPC/ML*

## AMD Instinct™ MI200 GPU SUPPORT

*20 key applications & frameworks on Infinity Hub & a catalogue supporting over 60 applications, frameworks & tools*

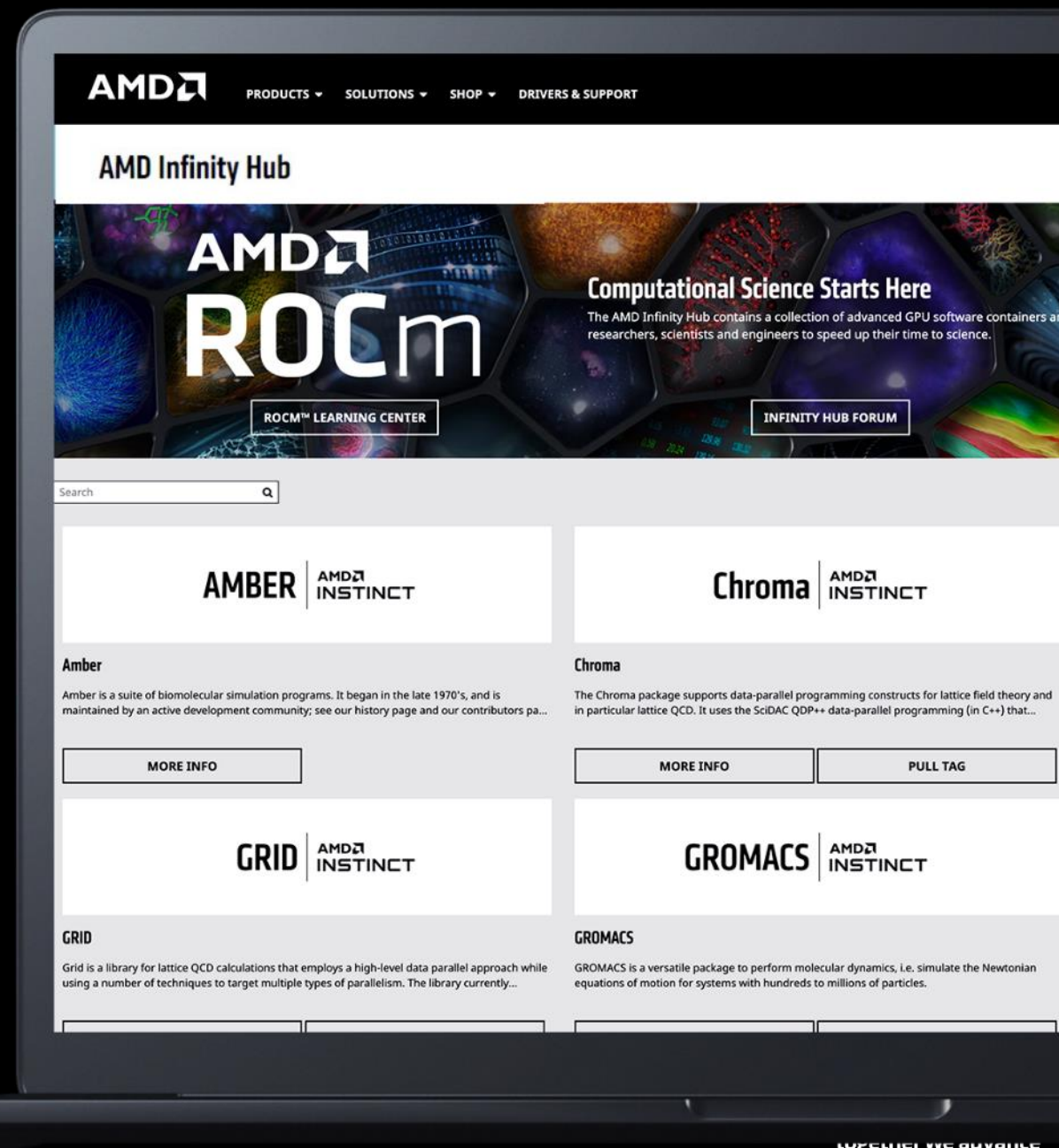
## Accelerating Instinct adoption

*Over 5000 application pulls since launch in the last year*

## PERFORMANCE RESULTS

*Published Performance Results for Select Apps / Benchmarks*

[AMD.com/InfinityHub](https://www.amd.com/InfinityHub)  
[Instinct Application Catalog](#)



# AMD INSTINCT™ COLLATERAL

## QUICK LINKS



AMD Instinct™  
MI250 Brochure



AMD CDNA™ 2  
White Paper



AMD Instinct™ Accelerator  
Qualified Servers



AMD Instinct™  
MI210 Brochure



AMD ROCm™ 5  
Brochure



GPU Accelerated  
Applications Catalog



# ***NEW ROCm™ TEXT BOOK***

AVAILABLE NOW



## Accelerated Computing with HIP

by Yifan Sun, Trinayan Baruah, David R Kaeli

Available through [Barnes and Noble](#)

ISBN-13: 9798218107444



# Thank you!

---