Introduction to ML Frameworks

Alessandro Fanfarillo, PhD
Frontier Center of Excellence
SMTS Software System Design Eng.
December 2022

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's current products, future products and markets, 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

01 CDNA and CDNA2 Architectures

02 Machine Learning, Deep Learning

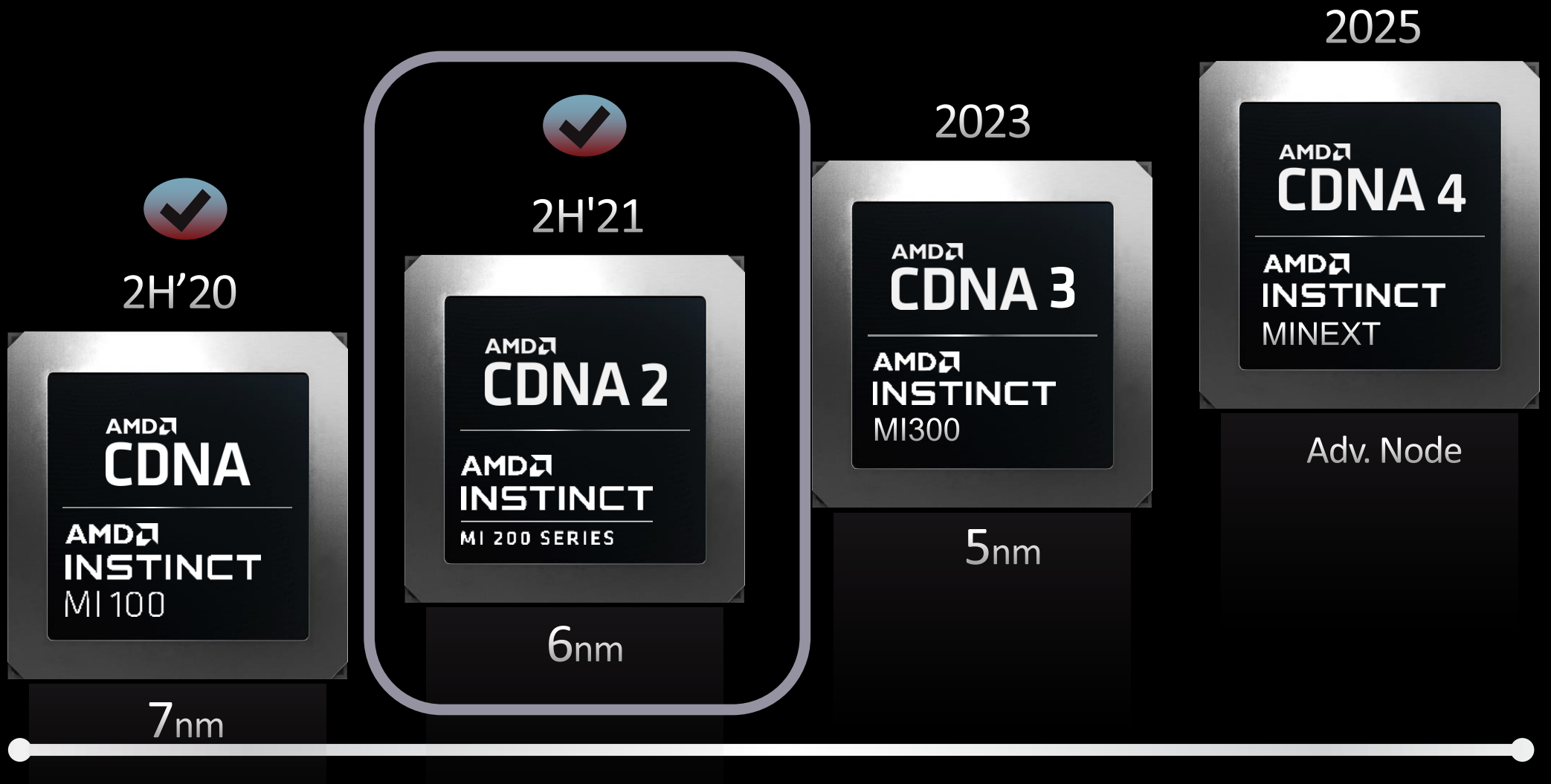
03 TensorFlow + Demo

04 PyTorch + Demo

01

CDNA/CDNA2 Architectures

COMPUTE GPU ARCHITECTURE ROADMAP

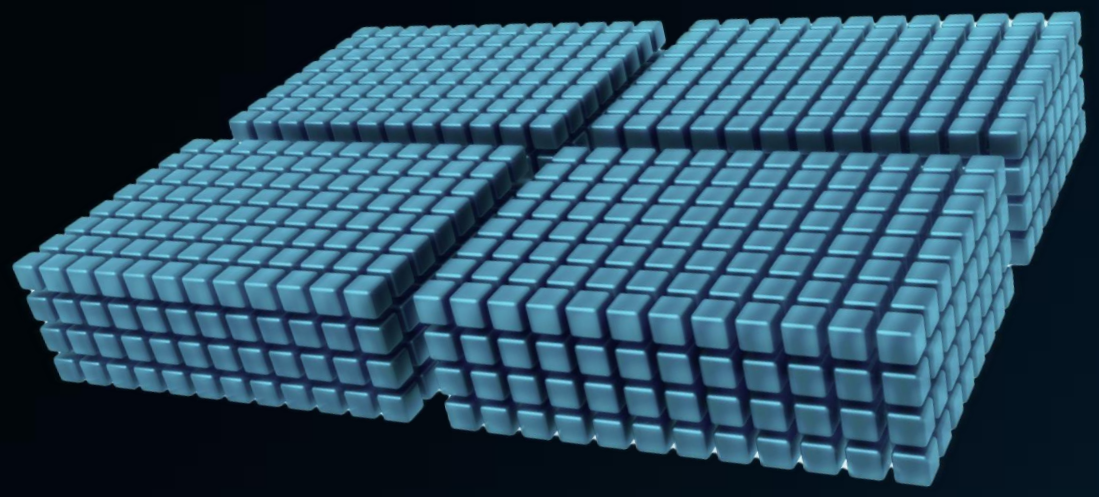


Roadmaps Subject to Change

AMD

2nd GENERATION MATRIX CORES

OPTIMIZED COMPUTE UNITS FOR SCIENTIFIC COMPUTING

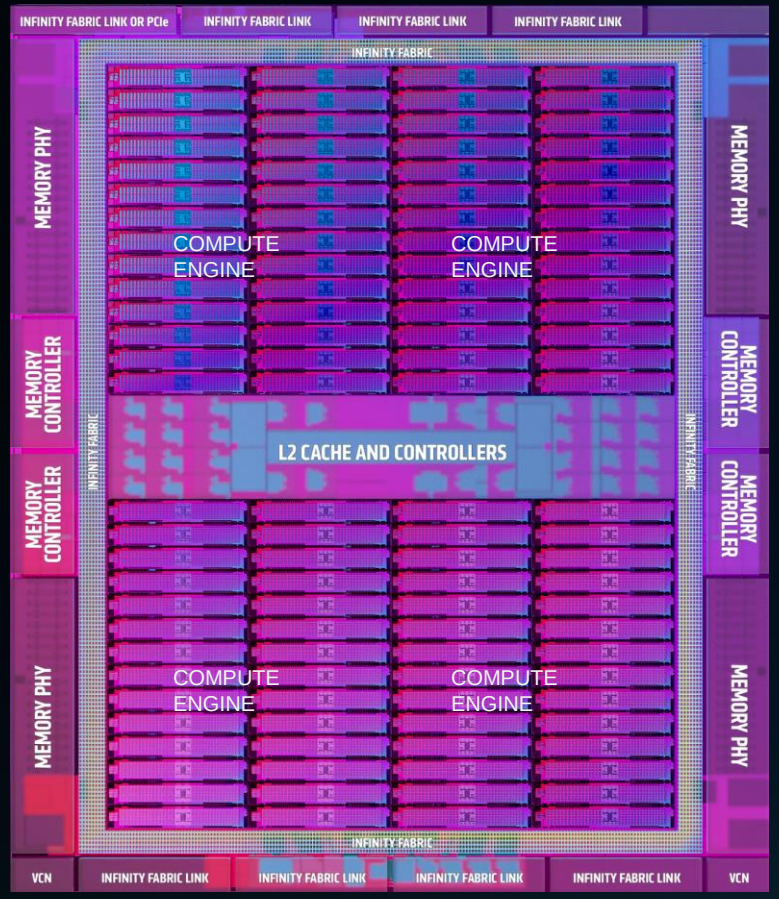


DOUBLE PRECISON (FP64)
MATRIX CORE THROUGHPUT
REPRESENTATION

MI100 MATRIX CORES OPS/CLOCK/COMPUTE UNIT	MI250X MATRIX CORES OPS/CLOCK/COMPUTE UNIT
No FP64 Matrix Core	256 FP64
256 FP32	256 FP32
1024 FP16	1024 FP16
512 BF16	1024 BF16
512 INT8	1024 INT8

2ND GENERATION CDNA ARCHITECTURE

TAILORED-BUILT FOR HPC & AI



- TSMC 6NM TECHNOLOGY
- UP TO 110 CU PER GRAPHICS CORE DIE
- 4 MATRIX CORES PER COMPUTE UNIT
- MATRIX CORES ENHANCED FOR HPC
- 8 INFINITY FABRIC LINKS PER DIE
- SPECIAL FP32 OPS FOR DOUBLE THROUGHPUT

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

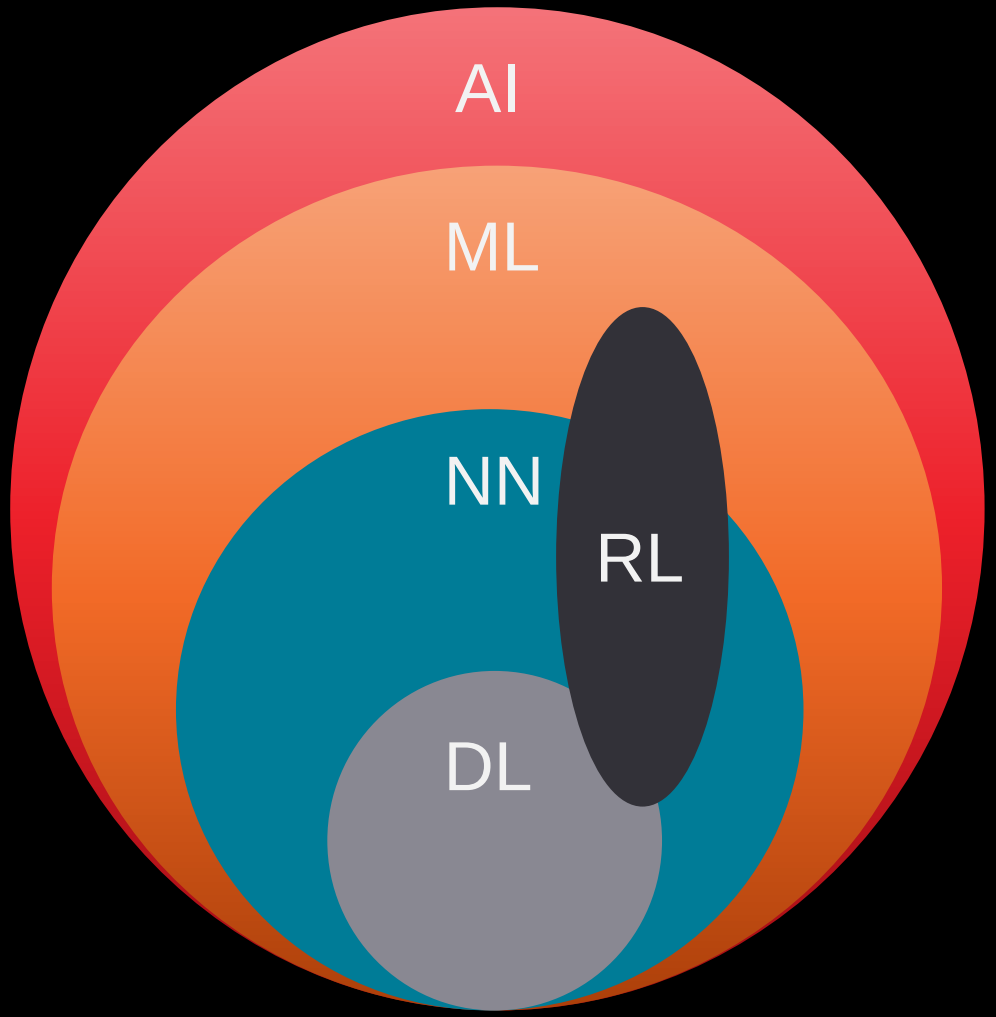
ENABLING DEVELOPER SUCCESS

- HPC Apps & ML Frameworks on AMD InfinityHub
- Streamlined and improved tools increasing productivity

02

Machine Learning, Deep Learning & AI

Machine Learning, Deep Learning, AI, RL, …



Artificial Intelligence (AI): techniques enabling computers to mimic human behavior

Machine Learning (ML): techniques that give computers the ability to learn without being explicitly programmed (learning from data)

Neural Networks (NN): models and techniques that teach computers to process data like the human brain (able to model non-linear distributions)

Deep Learning (DL): use of NNs with three or more layers (able to model very complex distributions)

Reinforcement Learning (RL): use models (ML/DL) to learn how to achieve a goal in unexplored/unpredictable environments (learning based on rewards and penalties)

AI/ML ECOSYSTEM SUPPORT & ADOPTION

UPSTREAMED SOURCE & BINARY SUPPORT

JAX | DEEPSPEED | CUPY

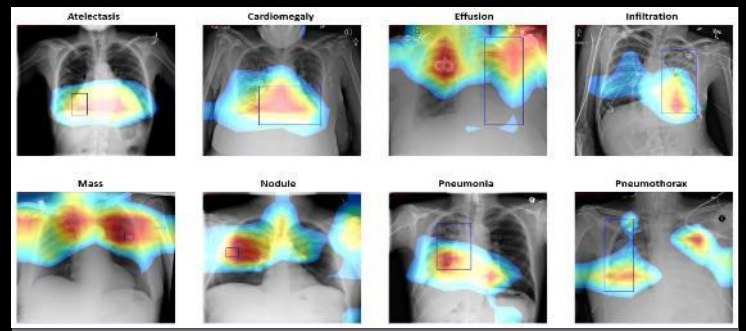
ML FRAMEWORKS & LIBRARIES

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

Focused on Targeted ML Use-Cases

Most common models on HuggingFace supported on AMD platforms today



VIDEO & IMAGE RECOGNITION

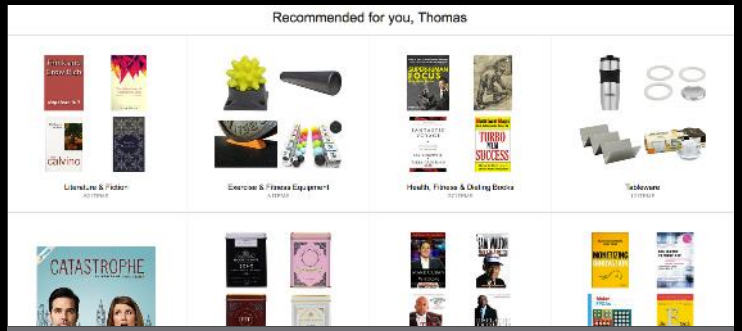
Optimized Models
Resnet, VGG, Inception
GoogleNet, ResNext, Detectron2,
RetinaNet, Mask R-CNN

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

AMD INFINITY HUB

Ready-to-Use HPC/ML Containers

AMD Instinct™ MI200 Support

Over 15 key applications & frameworks on Infinity Hub & a catalogue supporting over 50 applications, frameworks & tools

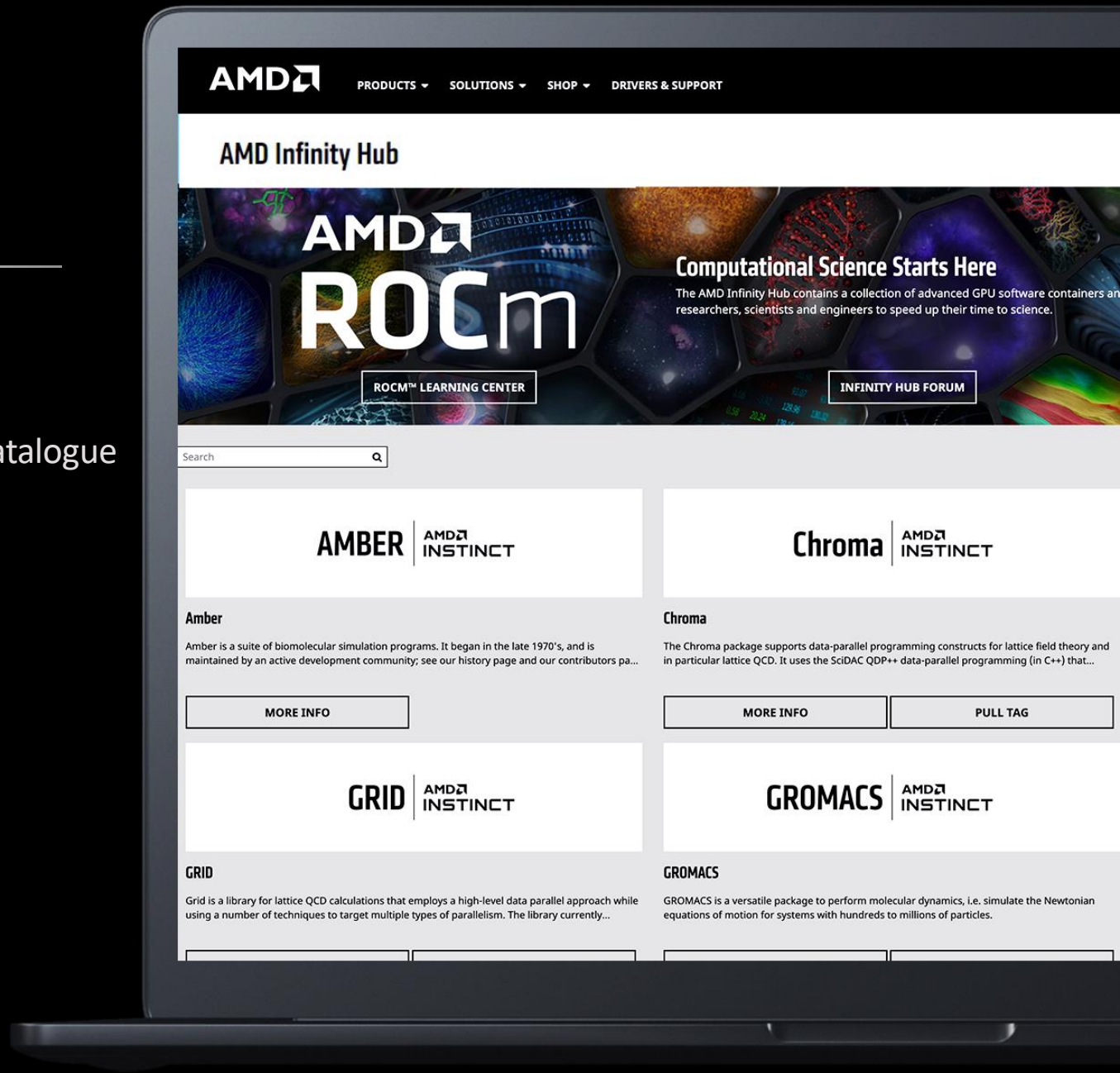
Accelerating Instinct™ Adoption

Over 5000 application pulls since launch last year

PERFORMANCE RESULTS

Published Performance Results for Select Apps / Benchmarks

AMD.com/InfinityHub
Instinct Application Catalog



AMD Instinct™ GPUs & ROCm™ SOFTWARE ECOSYSTEM USEFUL WEB RESOURCES

- AMD Instinct GPUs:
  - AMD Instinct™ MI100 GPU page: https://www.amd.com/en/products/server-accelerators/instinct-mi100
  - AMD Instinct™ MI210 GPU page: https://www.amd.com/en/products/server-accelerators/instinct-mi210
  - AMD Instinct™ MI Series Product Page: https://www.amd.com/en/graphics/instinct-server-accelerators
  - AMD Instinct™ HPC Solutions Page: https://www.amd.com/en/graphics/servers-instinct-mi-powered-servers
  - AMD Instinct™ Machine Learning Solutions Page: https://www.amd.com/en/graphics/servers-instinct-deep-learning
  - AMD CDNA2 Architecture: https://www.amd.com/en/technologies/cdna2
  - CDNA2 WP: https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf
- AMD ROCm™ open software platform:
  - AMD ROCm™ pages: https://www.amd.com/en/graphics/servers-solutions-rocm
  - AMD Infinity Hub: https://www.amd.com/en/technologies/infinity-hub
  - **AMD ROCm™ Deep Learning**: https://docs.amd.com/bundle/ROCm-Deep-Learning-Guide-v5.2/page/Deep_Learning_Training.html
  - ROCm Information Portal (DOCs & Learning Ctr.): https://docs.amd.com/
- HPC & AMD page: www.AMD.com/HPC

03

TensorFlow

TensorFlow and Keras

- TensorFlow (TF) is an open-source library for solving problems of ML/DL/AI
- Focus on training and inference in neural networks
- Functional and OOP capabilities
- Support for multiple accelerators (https://www.tensorflow.org/tutorials/distribute/keras)
- Keras is a high-level API for building NNs (adopted in 2017)
- Two options to install a ROCm-aware version of TF:
  1. Via docker image
  2. Via wheel package (Python 3.7+ is required)

For more info on how to install TF + ROCm please visit:

https://docs.amd.com/bundle/ROCm-Deep-learning-Guide_5.2/page/Frameworks_Installation.html

Tutorials available at: https://github.com/tensorflow/docs

More examples available at: https://github.com/tensorflow/examples

04

PyTorch

PyTorch

- PyTorch is an optimized tensor library for deep learning using GPUs and CPUs.
- Object oriented (python-like) and able to work on multiple accelerators
- Popularity increasing in the AI & research communities
- Two options to install a ROCm-aware version of PyTorch:
  1. Via docker image
  2. Via wheel package
  3. AMD is one of the founding members of the PyTorch foundation

For more info on how to install PyTorch + ROCm please visit:

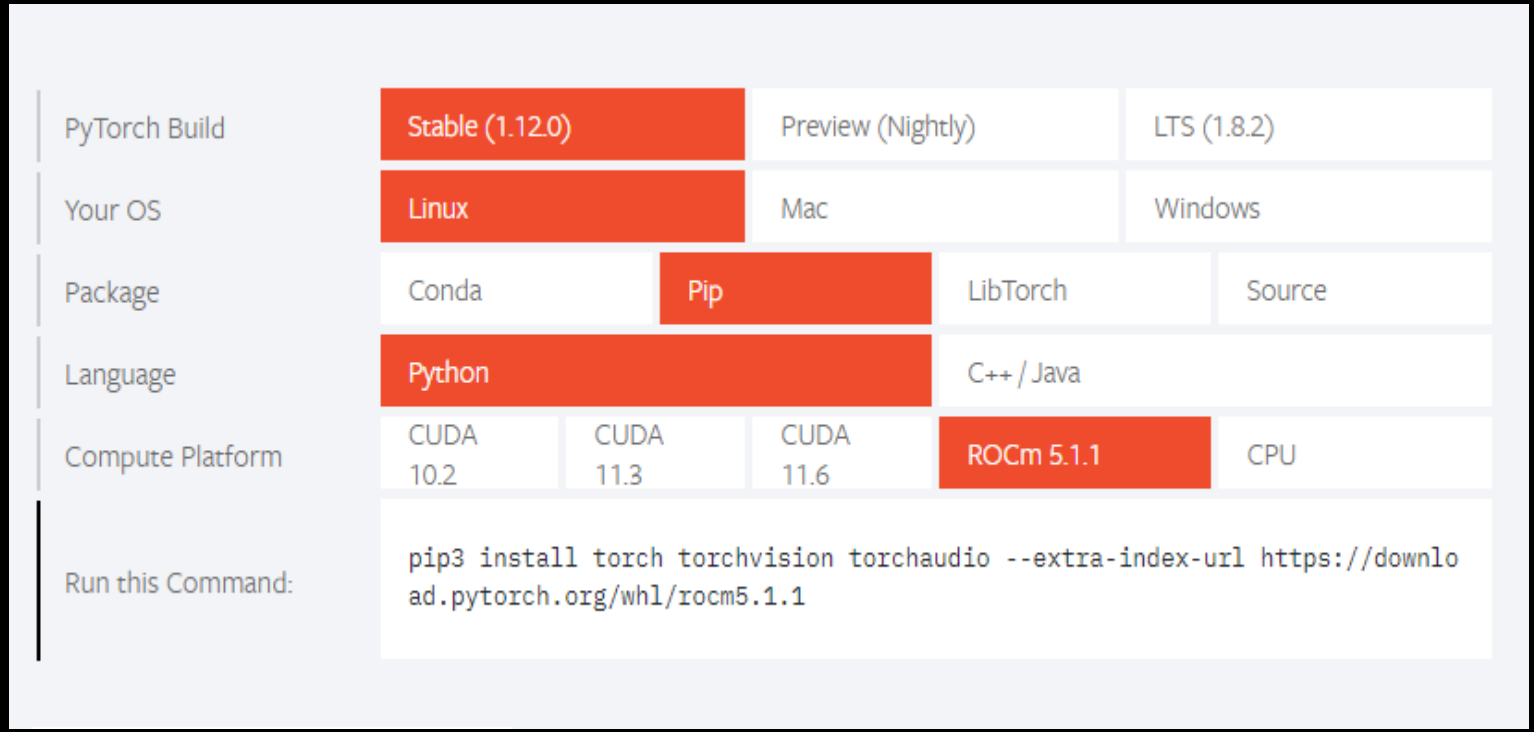
https://docs.amd.com/bundle/ROCm-Deep-learning-Guide_5.2/page/Frameworks_Installation.html

Tutorials available at : https://github.com/pytorch/tutorials

Examples available at: https://github.com/pytorch/examples.git

PyTorch 1.12.0

AMD ROCm™ SUPPORT THROUGH BINARIES FROM PyTorch.ORG



Disclaimer

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. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. 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.

THIS INFORMATION IS PROVIDED ‘AS IS.” 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 NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, 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.

Third-party content is licensed to you directly by the third party that owns the content and is not licensed to you by AMD. ALL LINKED THIRD-PARTY CONTENT IS PROVIDED “AS IS” WITHOUT A WARRANTY OF ANY KIND. USE OF SUCH THIRD-PARTY CONTENT IS DONE AT YOUR SOLE DISCRETION AND UNDER NO CIRCUMSTANCES WILL AMD BE LIABLE TO YOU FOR ANY THIRD-PARTY CONTENT. YOU ASSUME ALL RISK AND ARE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY ARISE FROM YOUR USE OF THIRD-PARTY CONTENT.

ANSYS, CFX and any and all ANSYS, Inc. brand, product, service and feature names, logos and slogans are registered trademarks or trademarks of ANSYS, Inc. or its subsidiaries in the United States or other countries.
Eclipse® is registered to the Eclipse Foundation in the United Stated, other countries, or both

The OpenMP name and the OpenMP logo are registered trademarks of the OpenMP Architecture

Questions?

alessandro.fanfarillo@amd.com


AMD