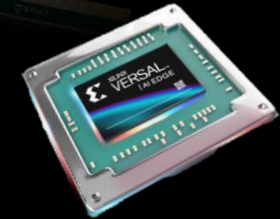
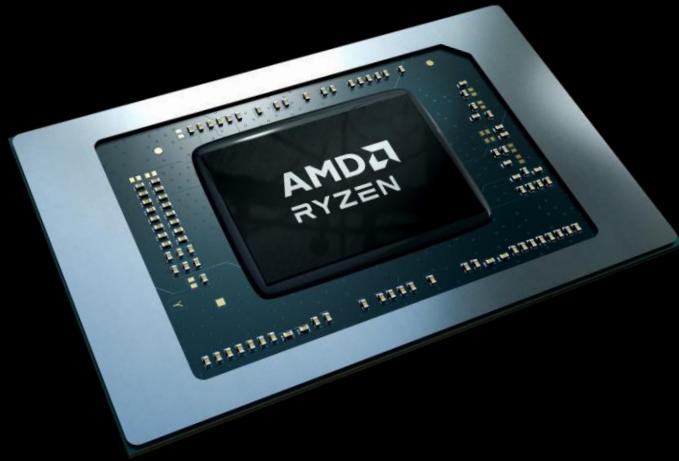


AMD accelerated compute update – Overview on latest Alveo and software road map




Jens Stapelfeldt – AI DBM & AI Eco-System Manager

Jens.Stapelfeldt@amd.com

LinkedIn: www.linkedin.com/in/JensStapelfeldt

AMD | Powers the daily lives of billions



Cloud



Healthcare



Industrial



Automotive



Connectivity



PCs



Gaming



AI

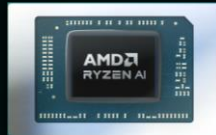


Announced today at **Computex 2024**

AI and high-performance leadership



Ryzen™ 9000 Series
Leadership gaming PCs
Available July



3rd Gen AMD Ryzen™ AI
Leadership AI PCs
Available July



Versal™ AI Edge Gen 2
Leadership AI Edge
Early access now



5th Gen AMD EPYC™
Best data center AI CPU
Available 2H 2024



AMD Instinct™ MI325X
Leadership AI accelerator
Available Q4 2024



Advancing end-to-end AI infrastructure

Cloud

HPC

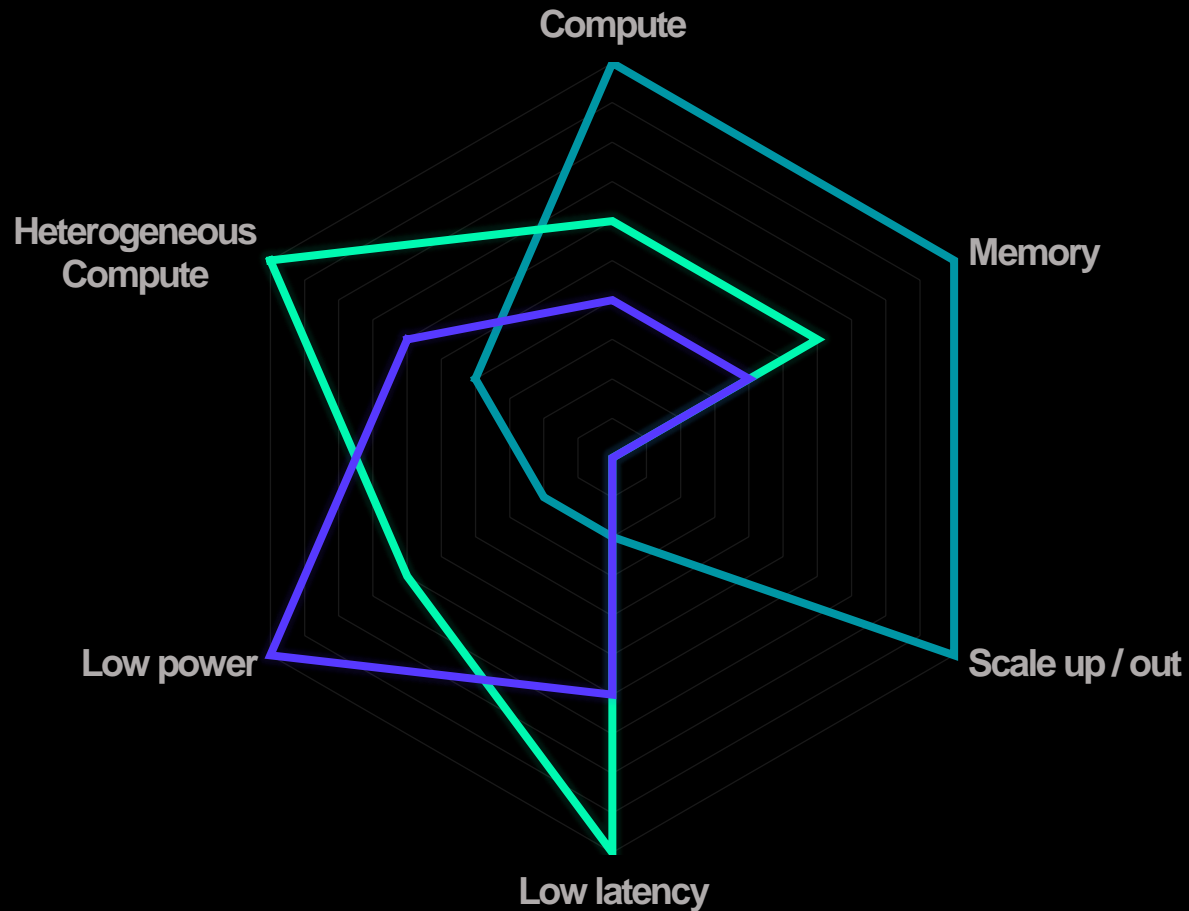
Enterprise

Embedded

PC

Diverse requirements

— Cloud AI — Edge / Embedded AI — Endpoint AI



Cloud / Enterprise

- 1000+ of TFLOPS compute
- TCO, demanding compute and memory
- Scale to thousands of nodes
- Diverse workloads

Edge / Embedded

- 10s-100s of TOPS
- “Hard” Real-time from sensor to control
- Whole application acceleration
- Form factor & safety focused workloads

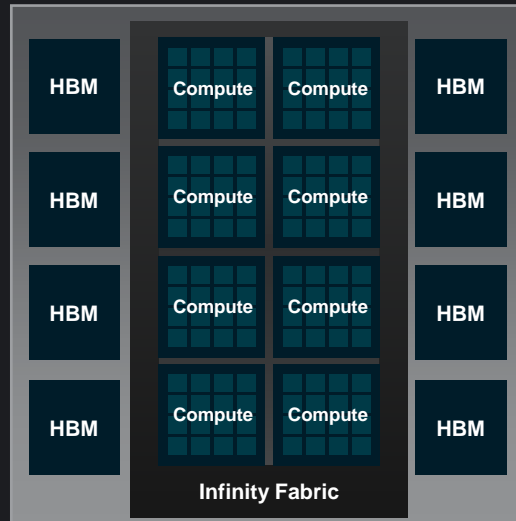
Endpoint

- 1-10s of TOPS with low precision support
- Performance per watt and battery focus
- Heterogeneous solution CPU-GPU-AI
- Privacy and data protected by on-device AI

Architectures to address diverse AI requirements



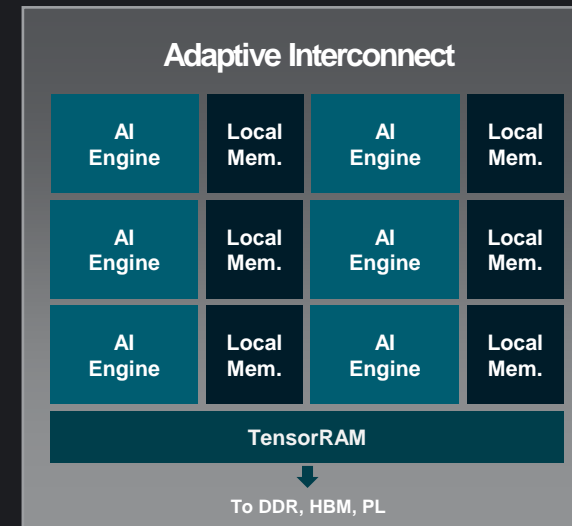
Data Center GPUs



- High performance GPU compute
- Large scale training and dense inference



Edge/Embedded and Client Endpoints



- Spatial data-flow compute
- Scalable and real-time inference

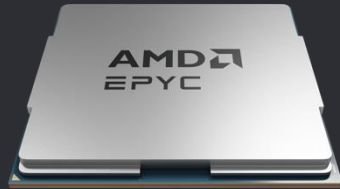
Broad AI compute portfolio



Cloud

Instinct™ MI300X

Data center training
and inference



Cloud and
Enterprise

EPYC™ Processors

CPU AI leadership



Gaming AI

Radeon™ W7900

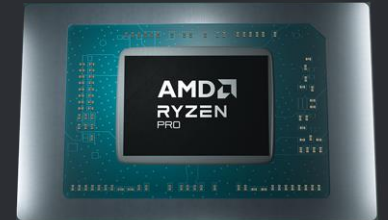
Gaming AI



Edge/Embedded

Versal™ AI Edge

AI + sensor
embedded inference

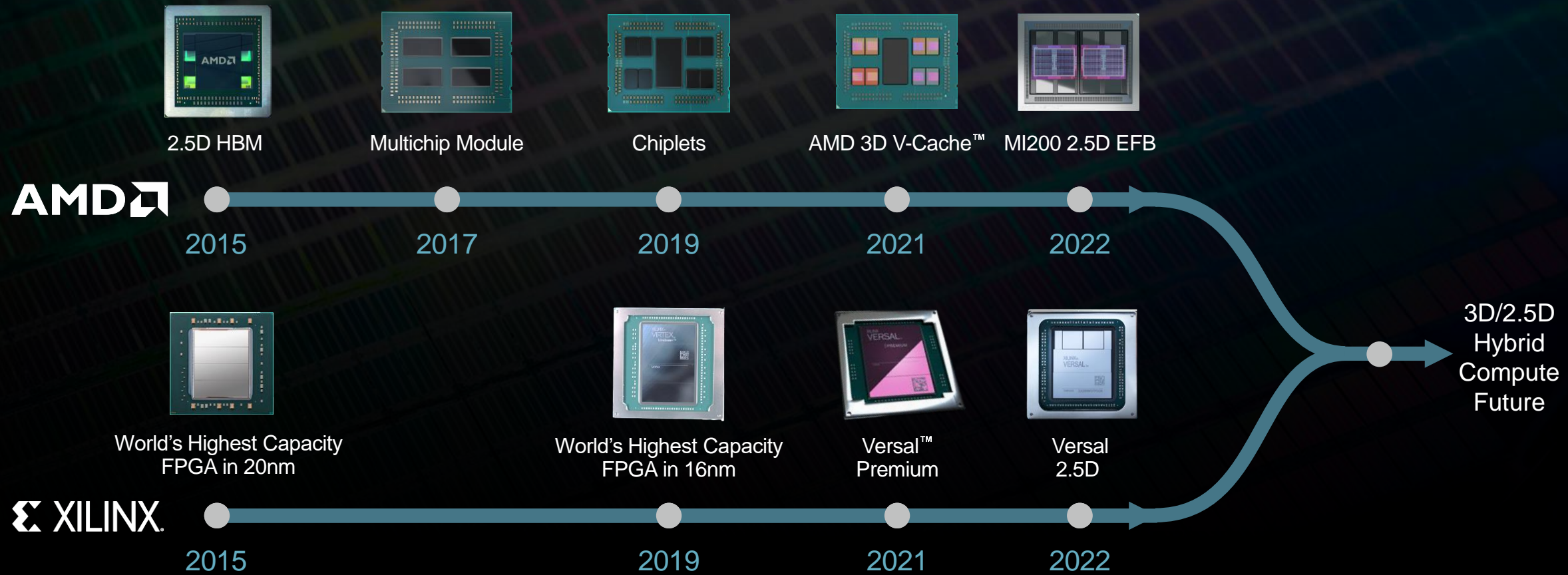


Endpoint

Ryzen™ AI 7040

AI inference for
Windows PCs

AMD Chiplet and Packaging Leadership



AMD Instinct™ MI300X

Leadership generative AI accelerator

Industry leading compute architecture with highest memory density & bandwidth

Unmatched TCO supporting up to 70B-parameter models* on a single GPU

2.5D / 3D chiplet construction with 153 B transistors

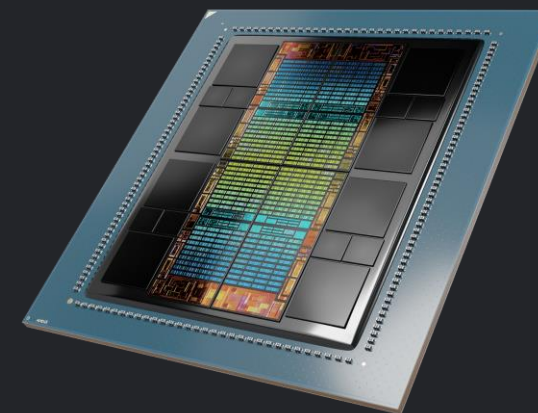
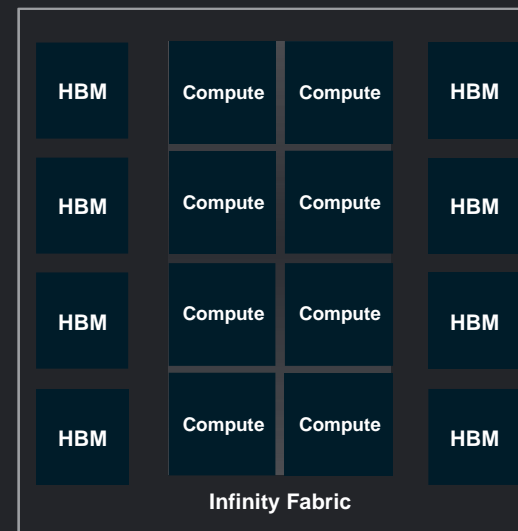
AMD
CDNA 3

192 GB
HBM3

5.2 TB/s
Memory Bandwidth

896 GB/s
Infinity Fabric™ Bandwidth

Data Center AI Architecture



Sampling now

*FP16 models

AMD Ryzen™ AI

The world's first dedicated AI engine on an x86 PC processor

10 TOPS (INT8) @ low single digit watts for maximum battery life

Spatial architecture for no-jitter, AI multi-task execution

Deployed in Windows Studio Pack and 3rd party applications across 35+ laptops



Up to 8 cores and
16 threads with
up to 5.2 GHz boost

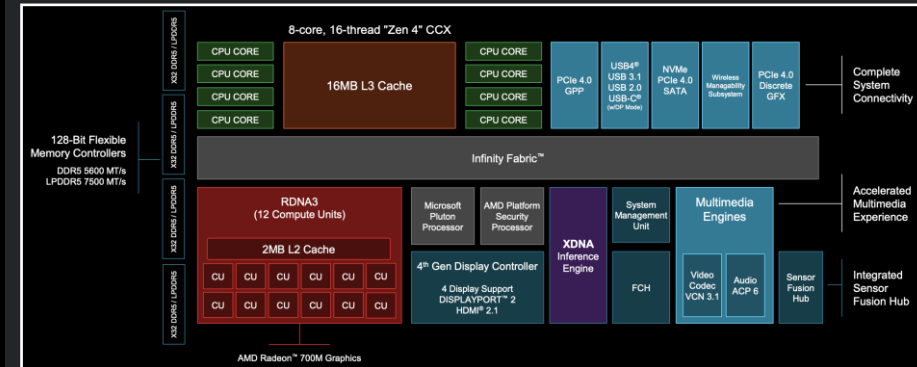
AMD
RDNA 3

AMD
XDNA

4nm
Process node

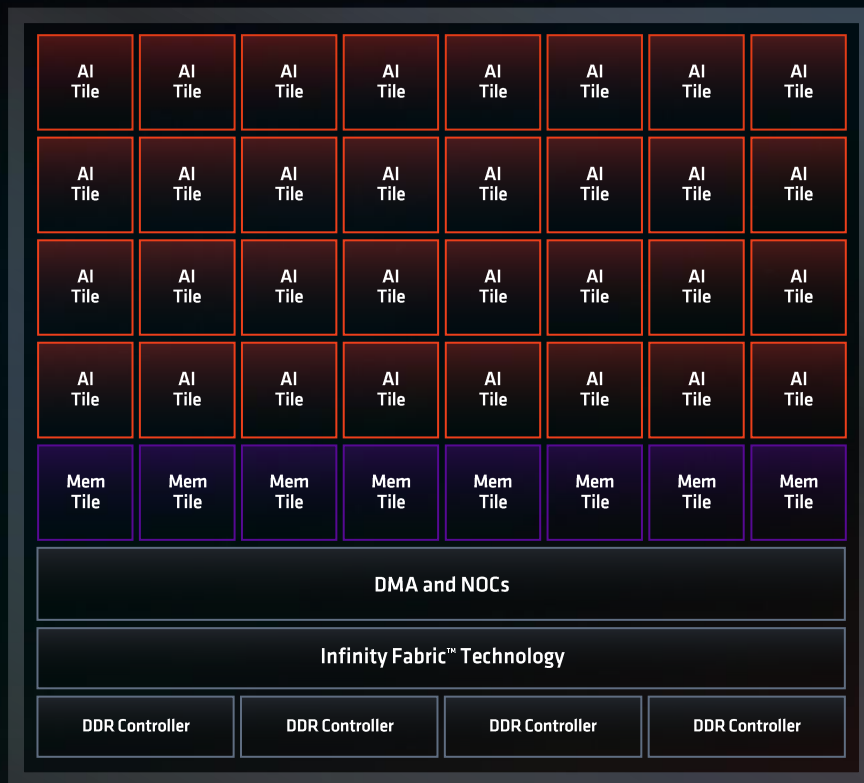
15-45 W
TDP

Endpoint AI Architecture





World's most powerful NPU for Copilot+ PCs



Up to
5x
vs. 7040 Series

Compute Capacity

Enhanced AI Tiles
enabling 2x multitasking

Up to
2x
vs. 8040 Series

Power Efficiency

Architectural advancements
for generative AI workloads

Generational Uplift vs. AMD XDNA

[Link to AI E details](#)

See endnote STX-13, STX-14

Versal™ AI Edge

Delivering breakthrough AI performance per watt for real-time systems

Scalable portfolio with AI engine, scalar engine & FPGA

Dataflow architecture for real-time end-to-end processing

Built on proven heritage of at scale deployments with functional safety in ADAS, industrial and healthcare space

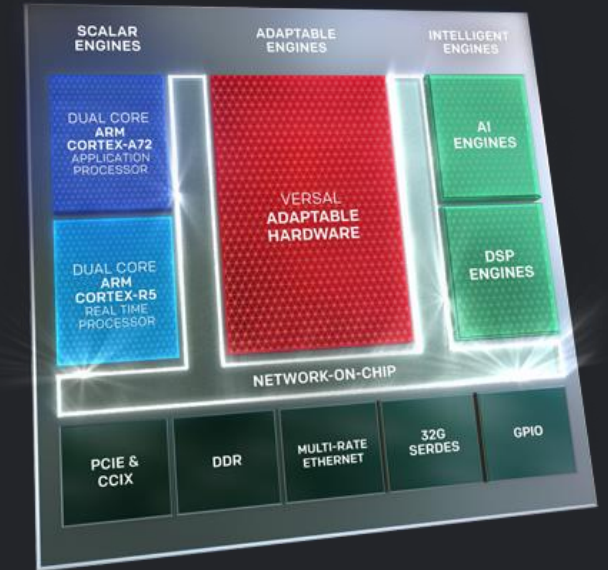
AMD
XDNA

10 - 400 TOPS
of AI compute

PCIe® 5.0

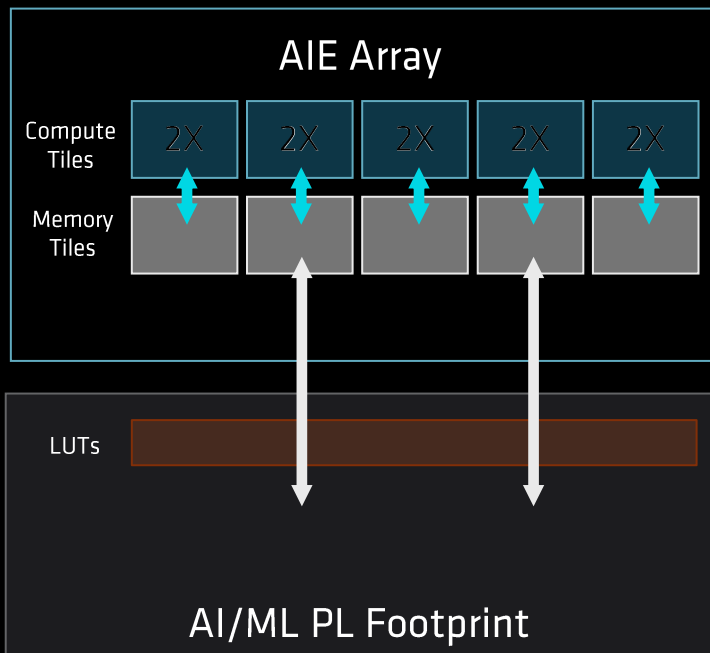
10-75 W
TDP

Edge / Embedded AI Architecture



AIE-ML v2: 2x Compute Per Tile With Expanded Data Type Support

AIE-ML Architecture



AIE-ML OPS / Tile

1024

INT4

512

INT8

256

BFLOAT16

No native support

FP16

No native support

FP8

No native support

MX9

No native support

MX6

AIE-ML v2

1024

1024

512

512

1024

1024

2048

Direct Memory BW / Tile / Cycle

512-bit

Load from
Local Memory

1024-bit

256-bit

Store from
Local Memory

512-bit

512-bit

Load from
Neighboring Memory

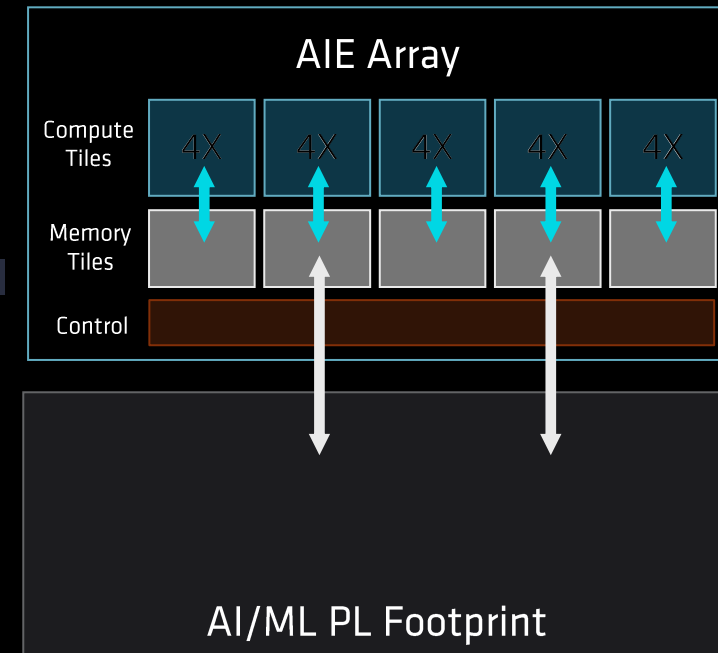
512-bit

256-bit

Store from
Local Memory

512-bit

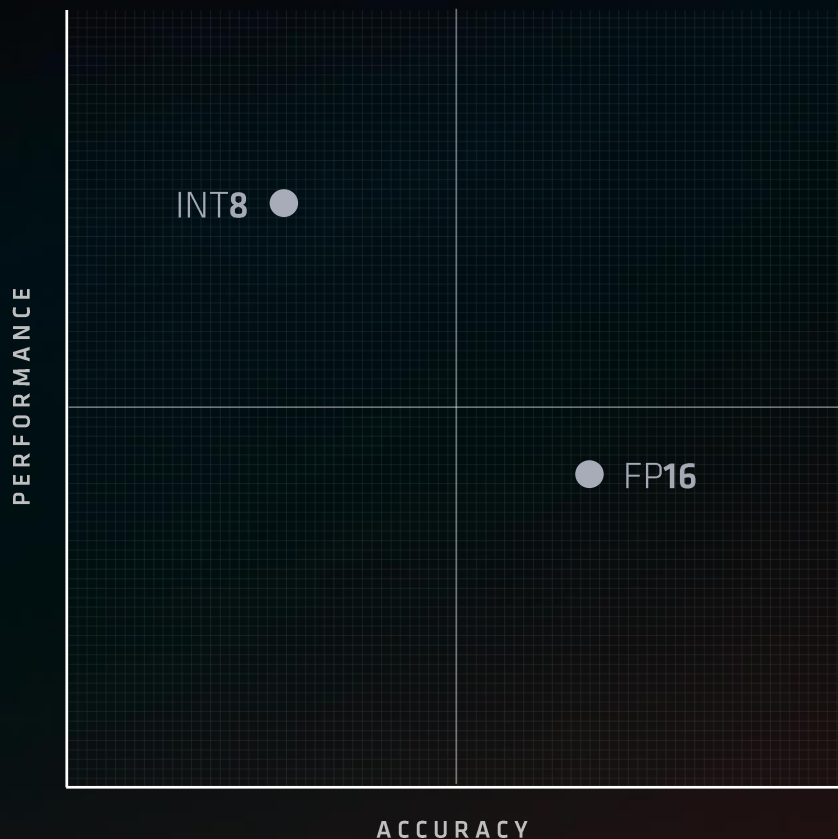
AIE-ML v2 Architecture



Datatypes details (MX)



Datatype is critical for advancing AI



Accuracy is critical for AI apps

8-bit

- Higher **Performance**
- Lower **Accuracy**

vs.

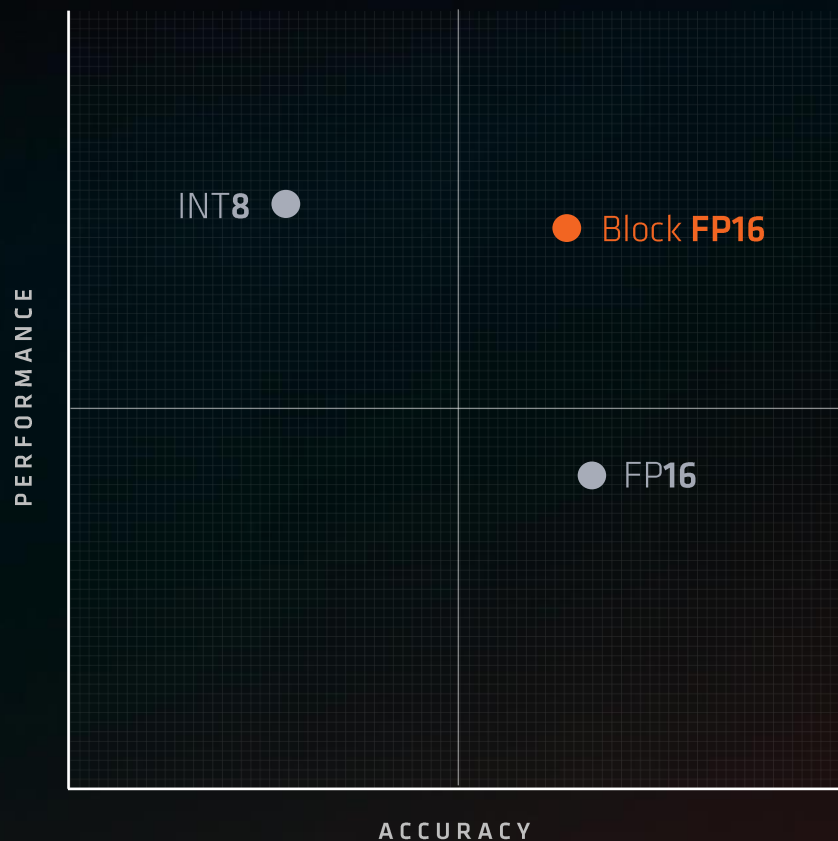
16-bit

- Higher **Accuracy**
- Lower **Performance**



3rd Gen AMD Ryzen™ AI

Datatype Leadership



Introducing world's first
Block Floating Point NPU

8-bit
PERFORMANCE

+

16-bit
ACCURACY

Majority of AI applications use 16-bit
No quantization needed

Alveo™ V70 COLLATERALS



Specification

Architecture	AMD XDNA™ – Versal AI Core
AI Engine	AIE-ML tiles
TOPS* (INT8 / BF16)	404 / 202
Memory Bandwidth	47.6 TB/s for internal memory 76.8 GB/s for DDR4 (16 GB)
Video Decoder**	96 channels of 1920x1080p
PCIe interface	Gen 4/5 x 8
Form Factor	Half Height, Half Length
Cooling	Passive
Power (TDP)	75 W

Part #	A-V70-P16G-ES3-G
SRP Pricing (1 unit)	\$1,995
Web Page	https://www.xilinx.com/applications/data-center/v70.html
Product Brief	https://www.xilinx.com/content/dam/xilinx/publications/product-briefs/alveo-v70-product-brief.pdf
V70 Software Lounge	https://www.xilinx.com/member/v70.html#overview
Online Store (qty <5)	https://www.xilinx.com/applications/data-center/v70.html#

* Using 50% weight sparsity

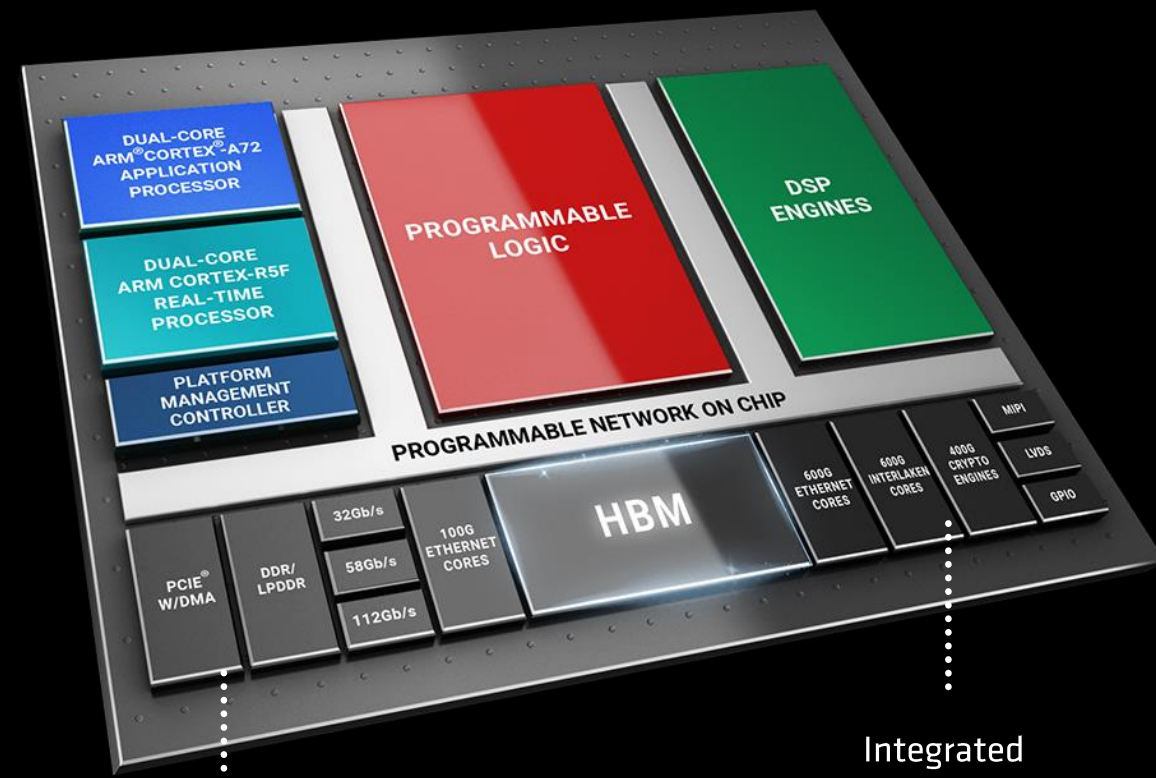
** 1920x1080p10 H.264/H.265

AMD Versal™ HBM Adaptive SoC Architecture

Powering the AMD Alveo™ V80 Accelerator Card

- Integrated, high-bandwidth networking cores and cryptographic engines
- 10,890 DSP slices for up to 2-3X greater DSP performance vs. previous generation
- Hardened connectivity to compute infrastructure for ease of integration

Conceptual Layout



Hardened Infrastructure Connectivity
(DDR Controller, PCIe® Gen5
w/DMA, Programmable NoC)

Integrated
High-Bandwidth Cores
(600G Ethernet, 400G Crypto)

DSP performance compared to AMD Alveo U55C card, see endnotes ALV-018 and ALV-019

AMD Alveo™ V80 Compute Accelerator Card Overview

7nm AMD Versal™ HBM Adaptive SoC

- 32GB HBM2e, 820GB/s bandwidth
- 2.6M LUTs of Programmable Logic

Memory Expansion

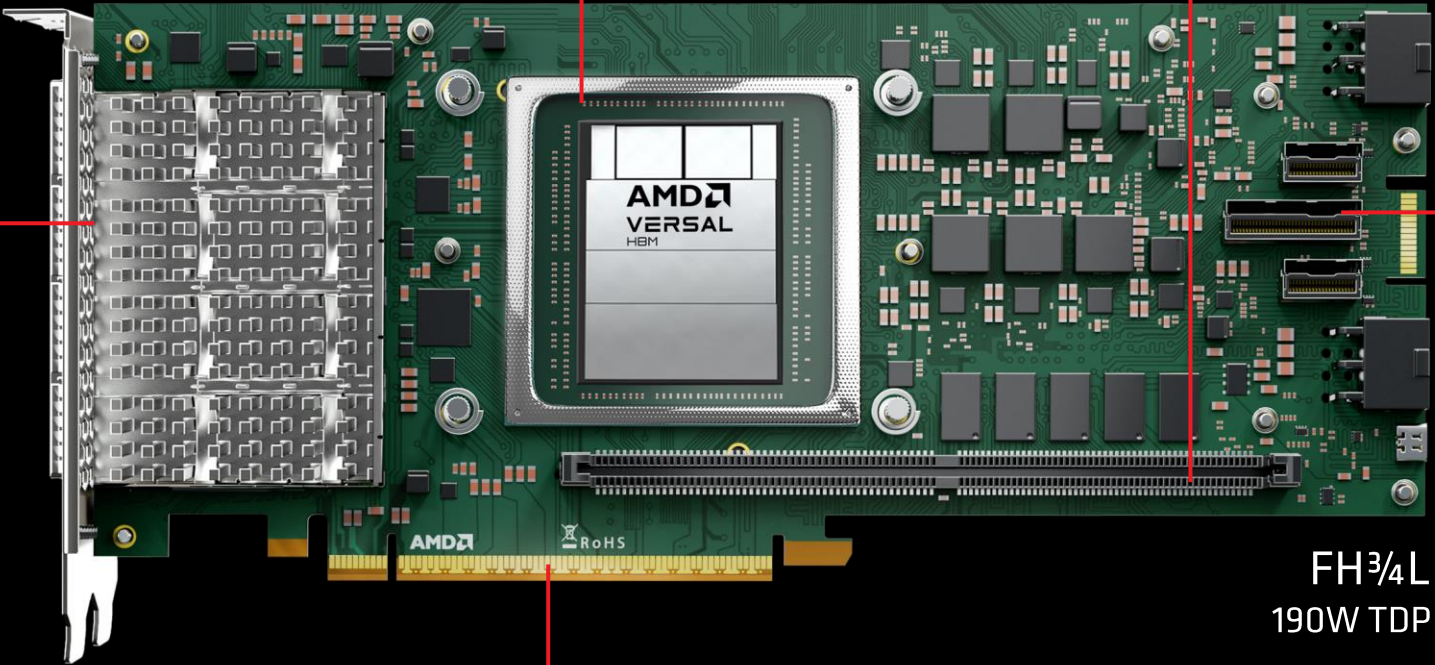
32GB DDR4 DIMM Expansion Slot

800G Bandwidth

- 4x200G or 4x 10/25/40/50G
- QSFP56 optical cages
- 58Gb/s Transceivers

MCIO Expansion Port

- Low latency 32Gb/s SerDes
- PCIe® Gen5
- Direct connectivity to NVMe drives
- Board-to-board for emulation

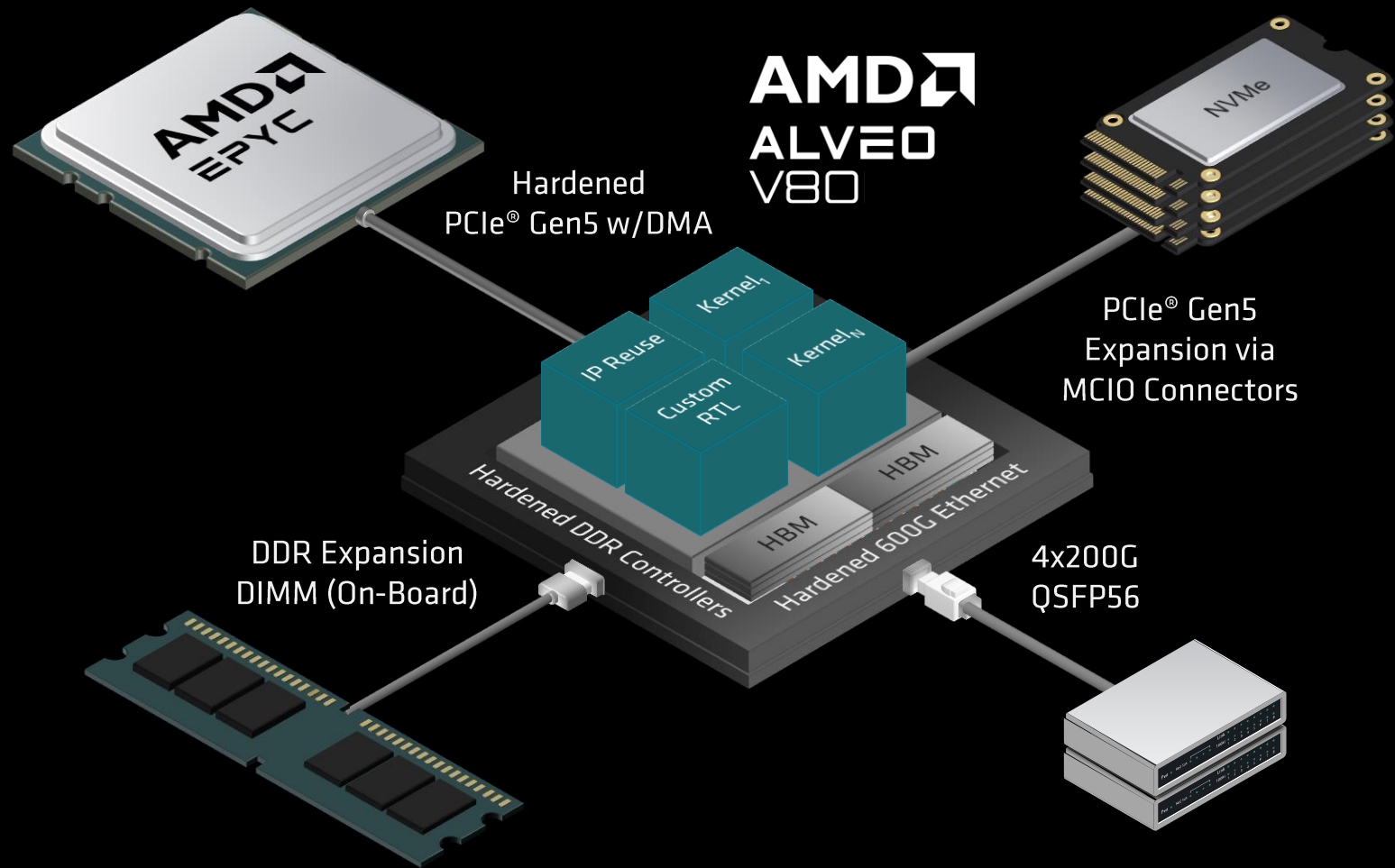


PCIe® Gen5

FH³/₄L
190W TDP

Passively cooled, total thermal power (TDP) is device and server dependent

Pre-Built, Hardened Connectivity to Data Center Infrastructure



Our compute accelerators are among the most versatile cards in the portfolio, targeting big data workloads with large data sets and those demanding massive parallelism, including data analytics, HPC, network switching, computational storage, blockchain, and various applications in FinTech including algorithmic trading and backtesting. The [Alveo U50](#) accelerator is our smallest form factor HBM-based card for memory bound compute, while the [Alveo U55C](#) card offers a balanced mix of HBM memory, logic density, and DSP resources in an FHHL form factor. The [Alveo V80](#) card is powered by the AMD Versal™ HBM adaptive SoC and delivers the highest logic density, HBM bandwidth, network bandwidth, and DSP resources in the Alveo™ portfolio—ideal for the most compute-intensive big data workloads.



Alveo U50



Alveo U55C



Alveo V80

Network Interface	1x100 Gb/s	2x 100 Gb/s	4x 200 Gb/s
Form Factor	HHHL	FHHL	FH ³ / ₄ L
Logic Density	872K LUTs	1,304K LUTs	2,574K LUTs
DSP Resources	5,952 Slices	9,024 Slices	10, 848 Slices
DDR4 Memory	-	-	32 GB
HBM Memory	8 GB	16 GB	32 GB
PCIe®	Gen4x8 or Gen3x16	Gen4x8 or Gen3x16	2x Gen5x8 or Gen4x16

Verify all data in this document with the device data sheets or product guides found at www.amd.com/alveo

Accelerator Card	Network Interface	Optical Interface	Form Factor	Logic Density	DSP	DDR4	HBM	AI Engines	PCIe®	Max Power ¹
Alveo U30	-	-	HHHL	N/A	N/A	8 GB	-	-	Gen3x8, 2x Gen3x4	75W
Alveo MA35D	-	-	HHHL	N/A	-	16 GB (LPDDR5)	-	-	Gen4x8	50W
Alveo U45N	2x 100G	2x QSFP28	FHHL	1M LUTs	1,320	8 GB	-	-	Gen4x8, Gen3x16	150W
Alveo U50	1x 100G	1x SFP28	HHHL	872K LUTs	5,952	-	8 GB	-	2x Gen4x8, Gen3x16	75W
Alveo U55C	2x 100G	2x QSFP28	FHHL	1.3M LUTs	9,024	-	16 GB	-	Gen3x16, 2x Gen4x8	150W
Alveo V80	4x 200G	4x QSFP56	FH¾L	2,574K LUTs	10,848	32 GB	32 GB	-	Gen4x16, 2x Gen5x8	300W
Alveo X3522PV	2x 10/25G	2x DSFP	HHHL	1M LUTs	1,320	8 GB	-	-	Gen4x8, Gen3x8	75W
Alveo UL3524	32x 10/25G	4x QSFPDD	FH¾L	787K LUTs	1,680	16 GB	-	-	Gen4x8	125W
Alveo V70	-	-	HHHL	N/A	N/A	16 GB	-	404 TOPs ²	Gen4x8, Gen5x8	75W

1. The maximum power represents the total electrical limit of the card. The thermal design power (TDP) of the Alveo card may vary depending on factors such as workload, inlet temperature, CFM/LFM airflow conditions, and the card's placement within the server. For specific TDP values, please refer to the associated Alveo accelerator card data sheet.

2. Alveo V70: INT8 Peak Performance with 50% Sparsity: <https://www.xilinx.com/applications/data-center/v70.html>

Our FinTech accelerators address the unique requirements of the financial market, which often demands a convergence of low-latency networking and hardware acceleration for algorithmic trading, pre-trade risk analysis, and market data delivery services at nanosecond speeds. The [Alveo™ X3522PV](#) and [Alveo UL3524](#) cards target low-latency and ultra-low latency trading, respectively, for algorithms that prioritize trade execution performance vs. algorithmic complexity. The [Alveo U55C](#) compute accelerators offer more resources for complex trading algorithms, risk & price modeling, as well as data analytics, while the [Alveo V80](#) card offers the most compute and bandwidth capability in the portfolio, ideal for financial modeling, analytics, and strategy backtesting.



Alveo X3522PV



Alveo UL3524



Alveo U55C



Alveo V80

Network Interface	4x 10/25G	32x 10/25G	2x 100 Gb/s	4x 200 Gb/s
Form Factor	HHHL	FH¾L	FHHL	FH¾L
Logic Density	1,030K LUTs	787K LUTs	1,304K LUTs	2,574K LUTs
DSP Resources	1,320 Slices	1,680 Slices	9,024 Slices	10, 848 Slices
AI Engines	—	—	—	—
DDR4 Memory	8 GB	16 GB	—	32 GB
HBM Memory	—	—	16 GB	32 GB
PCIe®	Gen4x8 or Gen3x8	Gen4x8	Gen4 x8 or Gen3 x16	2x Gen5x8 or Gen4x16
Expansion	—	32x 10/25G ARF6	—	2x Gen5x4, 1x Gen5x8 MCIO
Target Workloads	Low Latency Trading	Ultra-Low Latency Trading	Compute & Analytics	Algo Trading

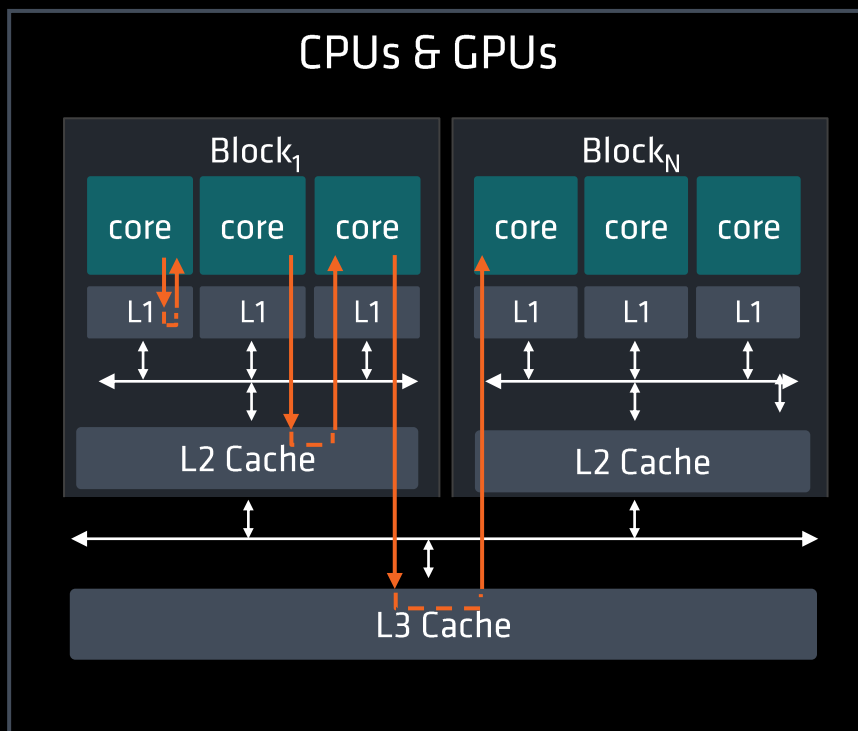
Verify all data in this document with the device data sheets or product guides found at www.amd.com/alveo

XMP451 (v2.1)

Adaptable Memory Hierarchy for Low Latency Processing

Traditional Architectures

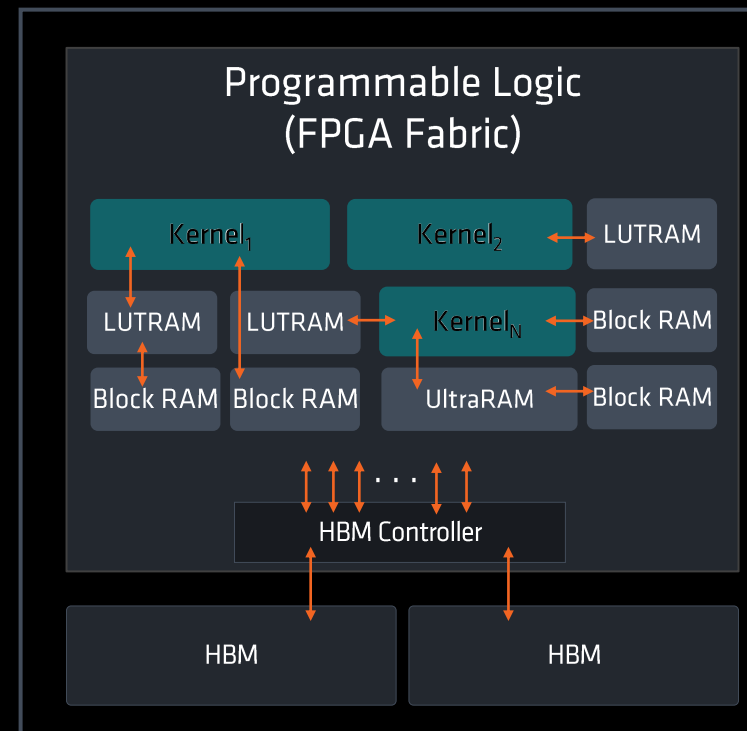
- Fixed cache hierarchy for data read/write
- Potential inefficiency for irregular access patterns



VS.

Adaptive Computing

- Allocate memory near compute for low latency & power
- Adaptable for custom data types & data movement



Adaptable Memory Hierarchy

- LUTRAM
- Block RAM
- UltraRAM
- HBM

Software is key to unlocking great hardware performance

Strengthening software capabilities

Enhancing developer experience

High-efficiency inference

Mipsology

Open-source AI compiler

nod.ai

Expanding our
open-source strategy

Advancing compiler-based
optimizations

Accelerating customer
engagements

Deploy Inference Models Rapidly with Vitis™ AI Development Environment

Complete AI Deployment Environment

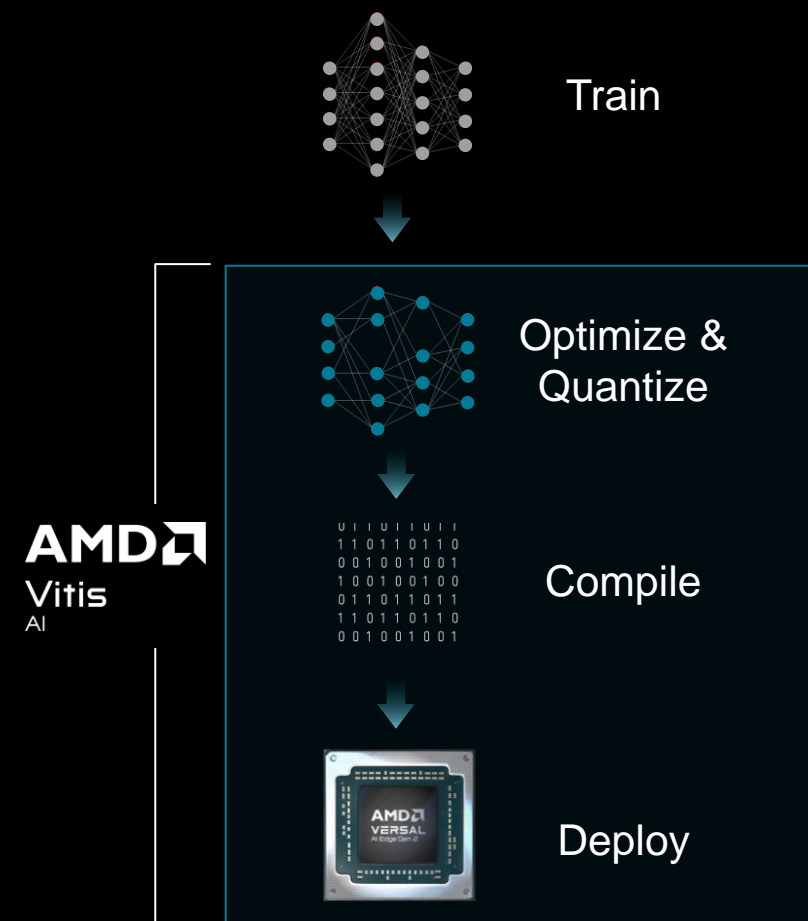
- Open-source with broad framework and model support

High-Performance IP

- Efficient implementation with no AI Engine coding

Advanced Optimizations

- Enhanced quantization algorithms optimize ease-of-use and accuracy
- Pruning to create sparse networks



Higher Levels of Programming Abstraction

AI Models and Algorithms



Ecosystem

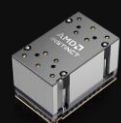
Libraries

Compilers and Tools

Runtime

AMD AI software

ROCm Vitis AI ZenDNN



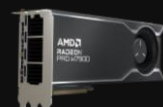
AMD Instinct™
Accelerators



AMD EPYC™
CPUs



AMD Alveo™
Accelerators



AMD Radeon™
Graphics



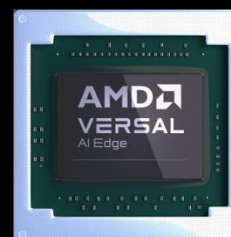
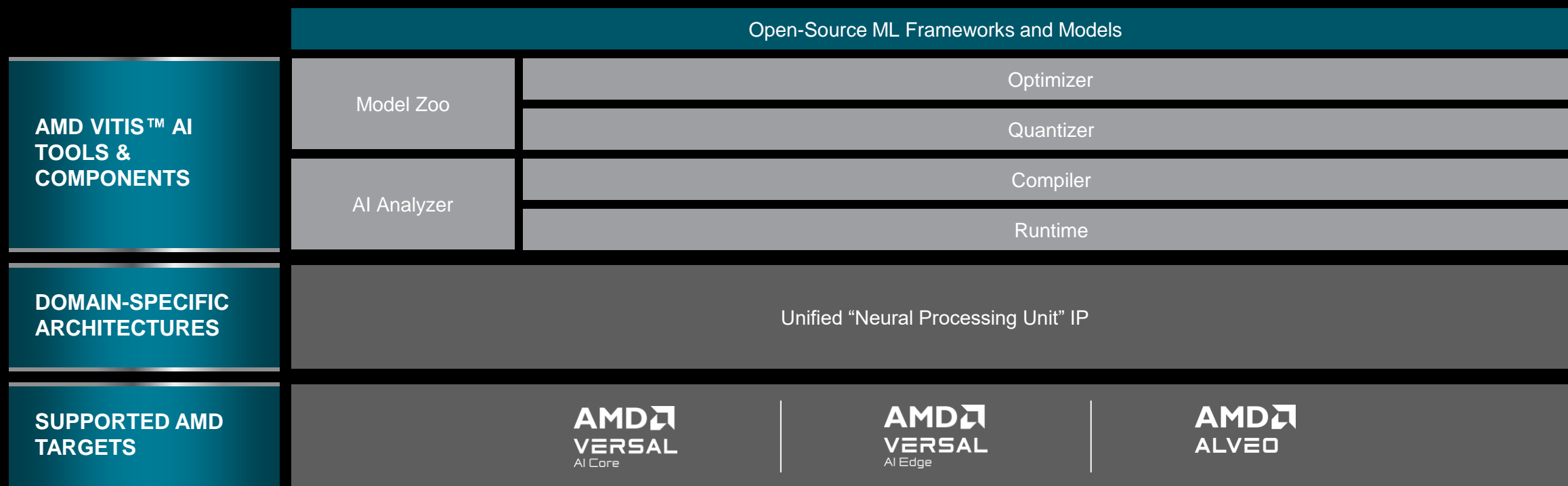
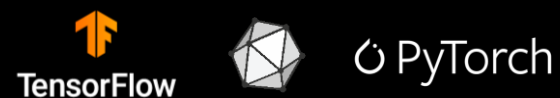
AMD Versal™
Adaptive SoCs



AMD Ryzen™
CPUs

AMD compute platforms

AMD AI Stack AIE-ML Support



VEK280

Strong Open Ecosystem Momentum



Hugging Face

62,000+ models running nightly
Fully integrated optimum library



PyTorch

From 'port-to' to 'develop-on'
with latest platforms



Tensor
Flow



Dynamo Inductor



JAX



OpenAI Triton



ONNX Runtime



OpenXLA



DeepSpeed



MLIR | IREE

Increasing open-source contributions
and expanding footprint

Q&A

LinkedIn: www.linkedin.com/in/JensStapelfeldt



Jens Stapelfeldt

Thought Leader, Tech. Lead, AI, Machine Learning, Data Center, Robotics, AI Vision, Lin...



ACCELERATING AI AND SCIENCE

AMD 
HPC Fund

AI & High Performance
Compute Fund

20+

PETAFLUPS

COMPUTING POWER

combined that would
rank among the **fastest**
supercomputers in the world*

AMD
HPC Fund

\$31M

**TOTAL
MARKET
VALUE**

28 ▶ 9

GRANTEES

COUNTRIES

US • CANADA • FRANCE • GERMANY
ITALY • SWITZERLAND • INDIA • SINGAPORE

400+

CPUs

AMD
EPYC

2800+

GPUs

AMD
INSTINCT






















100+

Alveo Cards

AMD
ALVEO

*According to the most recent Top 500 list

AI & HPC Fund Grantees

 MIT Massachusetts Institute of Technology	 Stanford University	 NYU	 TEXAS The University of Texas at Austin	 UNIVERSITY OF TORONTO	UCLA	 RICE Unconventional Wisdom
Carnegie Mellon University		 Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften	 NUS National University of Singapore	HLRS High-Performance Computing Center Stuttgart	 HOWARD UNIVERSITY 1867	 CANCER RESEARCH UK CAMBRIDGE CENTRE
 The University of Vermont	 VCU School of Pharmacy	 UNIVERSITÀ DI TRENTO Department of Information Engineering and Computer Science	TEXAS★STATE UNIVERSITY <i>The rising STAR of Texas</i>	Oregon State University	 NORTH CAROLINA AGRICULTURAL AND TECHNICAL STATE UNIVERSITY	
 Washington University in St. Louis	 GENCI Le calcul intensif au service de la connaissance	 INES Centre Informatique National de l'Enseignement Supérieur	 UNIVERSITY OF ARKANSAS	 Boston Children's Hospital	 THE UNIVERSITY OF BRITISH COLUMBIA	

<https://www.amd.com/en/corporate/hpc-fund>

AI & HPC Fund Research Cluster

- Intends to put significant computational power in the hands of those conducting high-impact AI and HPC research
- Open for applications from academics and lab researchers (typically with a 1-year allocation)
- Provides access to AMD EPYC series processors and Instinct™ MI series accelerators and associated programming environment



AMD

ROCm

PyTorch

TensorFlow

AI & HPC Fund Research Cluster – Software Environment

- Standard modules environment (Lmod) organizes pre-installed software (e.g., ROCm, HIP compilers, MPI library, development tools)
- ROCm-enabled PyTorch and TensorFlow provided via
 - Python environments
 - Singularity containers
- Jupyter Notebooks/Labs

Jupyter notebooks/labs supported through Slurm

```

----- /opt/ohpc/pub/moduledeps/gnu12-openmpi4 -----
omb/7.0.1

----- /opt/ohpc/pub/moduledeps/gnu12 -----
hdf5/1.10.9      openmpi4/4.1.5 (L)

----- /opt/ohpc/pub/modulefiles -----
autotools      (L)    hpcfund      (L)    prun/2.3      (L)    ucx/1.14.0    (L)
cmake/3.25.2    os              rocm/5.4.2
gnu12/12.2.0 (L)    pmix/4.2.2      rocm/5.5.1 (L,D)
valgrind/3.20.0

----- /share/modulefiles -----
omniiperf/1.0.7  pytorch/2.0.1  tensorflow/2.11.0
  
```

The screenshot shows a Jupyter Notebook titled 'Untitled2' with the following content:

```

In [1]: !rocm-smi

===== ROCm System Management Interface =====
===== Concise Info =====
GPU  Temp (DieEdge) AvgPwr  SCLK  MCLK  Fan  Perf  PwrCap  VRAM%  GPU%
0    35.0c        33.0W   300Mhz 1200Mhz 0%  auto  290.0W  0%  0%
1    35.0c        31.0W   300Mhz 1200Mhz 0%  auto  290.0W  0%  0%
2    35.0c        33.0W   300Mhz 1200Mhz 0%  auto  290.0W  0%  0%
3    34.0c        35.0W   300Mhz 1200Mhz 0%  auto  290.0W  0%  0%
===== End of ROCm SMI Log =====

In [2]: !import torch

In [3]: !print("GPU(s) available:", torch.cuda.is_available())
GPU(s) available: True

In [4]: !print("Number of available GPUs:", torch.cuda.device_count())
Number of available GPUs: 4

In [5]: !hostname
t006-004.hpcfund

In [ ]: !
  
```


AI & HPC Fund Research Cluster - Infrastructure

Provides a familiar HPC Linux cluster environment:

- Dedicated node access through resource manager (SLURM)
- Interactive login node for development
- High-speed interconnect with GPU aware MPI
- Persistent file storage on parallel file system(s)
- Containerization support
- Python productivity tools

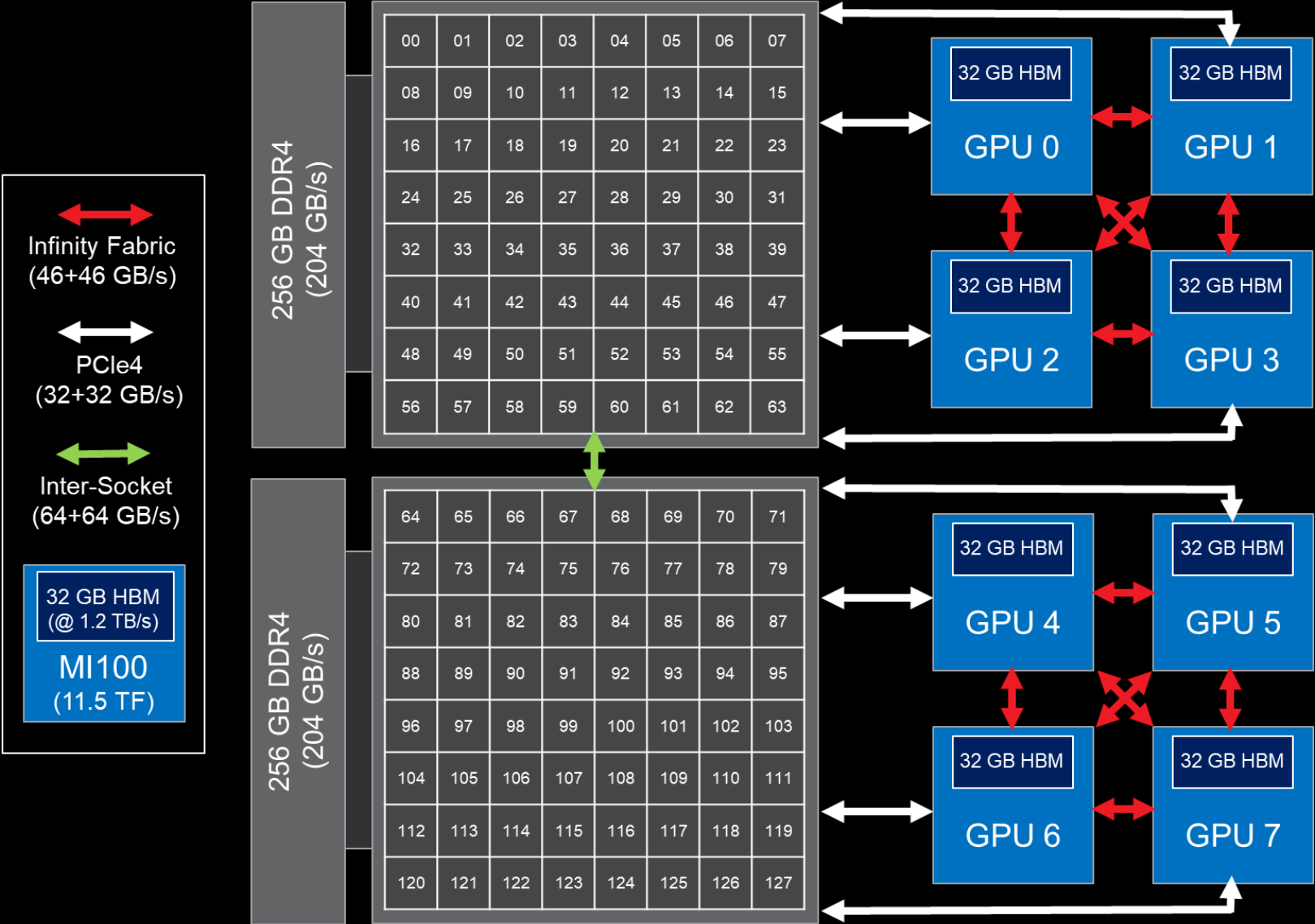
Compute Element(s)

64-core EPYC 7V13 CPU
MI100 or MI210 GPUs
ConnectX®-6 HCAs

40 total nodes with both 4
and 8 GPUs per node



AI & HPC Fund Research Cluster – Compute Node Diagram



AI & HPC Fund Research Cluster – How to Request Access

Application Form

<https://www.amd.com/en/forms/registration/amd-hpc-fund-research-accelerator.html>

Research applications in AI/ML are particularly encouraged!

AMD HPC Fund Application

Submit your application and we'll be in touch.

Please use this application form to provide us with information on your research project and to be considered for cloud access grants to AMD leading high-performance computing technologies, customized technical training as well as the opportunity to network with your peers around the world.

Once your application form is received by the AMD team, it will be reviewed quarterly (as mentioned in the [website FAQ](#)). With a finite number of grants every quarter we strive to support as many research organizations and academic institutions as possible to meet their project goals. However, if your application is not selected in the quarter you applied, your submissions will automatically be included for consideration in the next quarter.

We are excited about this opportunity for your organization and look forward to learning more about your project(s).

APPLICATIONS RECEIVED BY

MARCH 1

JUNE 1

SEPTEMBER 1

NOVEMBER 15

DECISION NOTIFICATION

APRIL 1

JULY 1

OCTOBER 1

JANUARY 1



User Support

<https://github.com/AMDRResearch/hpcfund>

Search or jump to...

Pulls Issues Codespaces Marketplace Explore

AMDResearch / hpcfund Public

Edit Pins Unwatch 3 Fork 1 Star 1

User Guide

<https://amdresearch.github.io/hpcfund>

AMD

Search docs

HPC FUND RESEARCH CLOUD

User Guide

View page source

User Guide



Heterogeneous Accelerated Compute Clusters

Enabling Novel Research in Heterogeneous
Compute Acceleration for HPC

Heterogenous Accelerated Compute Clusters (HACCs)

- A special initiative to support novel research in heterogenous compute acceleration for AI and HPC
- The scope encompasses systems, architecture, tools and applications
- Established at six of world's most prestigious universities
- Each is equipped with the latest AMD hardware and software technologies
- The goal is to foster a community of leading academic teams to conduct state-of-the-art research

Research Areas

- **AI and Machine Learning**
- Adaptive Compute Acceleration
- High Performance Computing (HPC)
- Database Acceleration
- Energy Efficiency
- Compilers
- Computer architecture (heterogenous computing systems)

HACCs: Heterogeneous Accelerated Compute Clusters

- Remote access to Adaptive Compute hardware
- HACC user group meetings
- Access to AMD researchers
- Collaboration opportunities



AMD
EPYC

AMD
INSTINCT

AMD
ALVEO

AMD
VERSAL

www.amd-haccs.io

★ Newest HACC at IISc, Bangalore

HACC Adaptive Computing Hardware



- HACC hardware consists of:
 - Compute and Alveo nodes (initially U250 and U280 with HBM)
 - Latest heterogeneous nodes (SMC 4124GS) include:
 - 2 EPYC Milan/X CPUs
 - 4 MI210 GPUs
 - 2 Alveo U55C FPGA with HBM
 - 2 VCK-5000 ACAP/Versal with AIEs
 - Run-time via ROCm, XRT
 - SW development via HIP, Vitis, frameworks
 - 100G network
- Community hub for researchers
 - Support from in-house AMD research groups
 - Reproducible results & experiments



