The Journey to AI Pervasiveness

Victor Peng

President, AMD

AMD

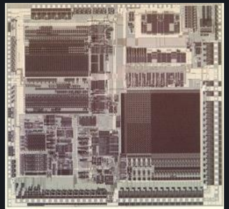
together we advance_

Evolution of AI

Early History: Supervised Learning, Perceptrons, MNIST

1960s

1989



V11 M Chip

~1984

ISSCC94 / SESSION 13 / NEURAL NETWORKS AND IMAGE SENSORS / PAPER TP 13.1

TP 13.1: A 1.2GFLOPS Neural Network Chip Exhibiting Fast Convergence

[illegible]

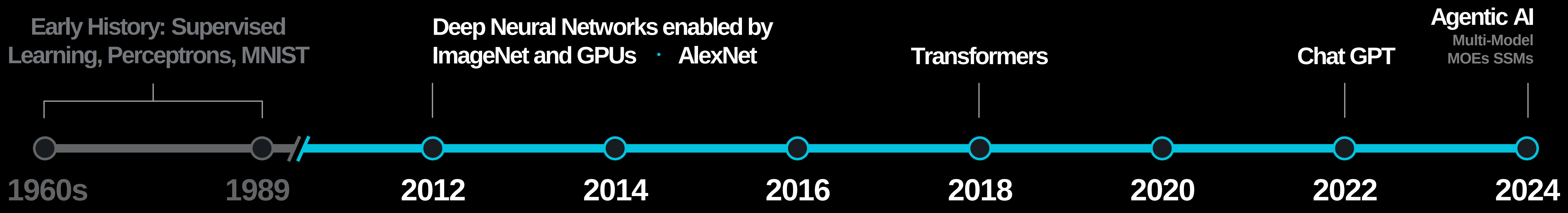
Early NPUs at ISSCC

1991

Alpha 21264

~1998

Evolution of AI



Early History: Supervised Learning, Perceptrons, MNIST

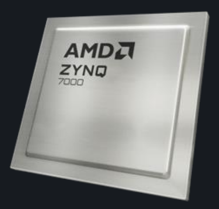
Deep Neural Networks enabled by ImageNet and GPUs · AlexNet

Transformers

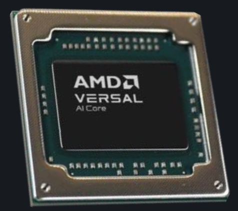
Chat GPT

Agentic AI
Multi-Model
MOEs SSMs

1960s · 1989 · 2012 · 2014 · 2016 · 2018 · 2020 · 2022 · 2024

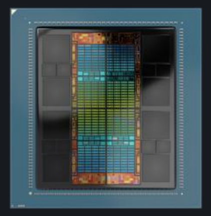


First FPGA SOC
AMD Zynq™ 7040
~2013



First AIE Device
VC1902
~2021

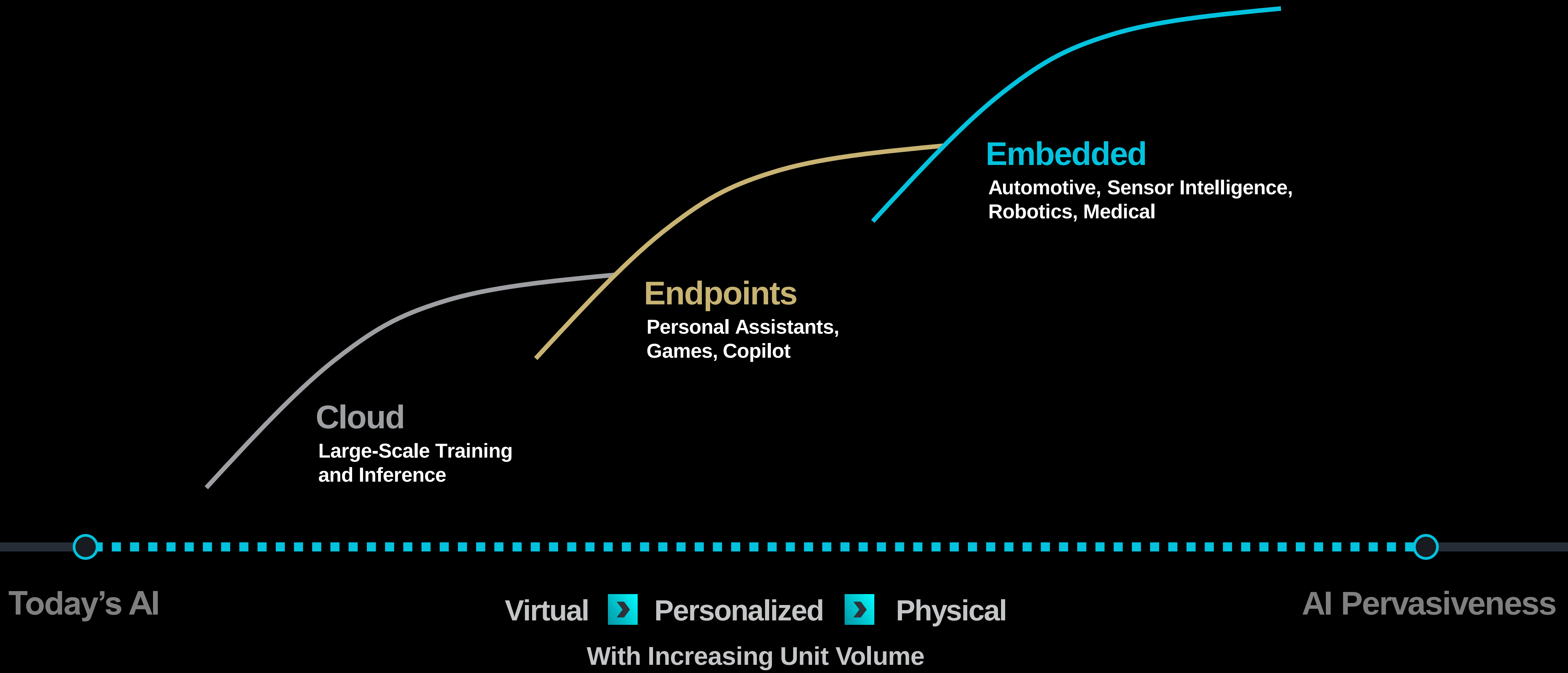
Largest FPGA Ever
VP1902
~2022



Leadership GPU for AI Inference
MI300X
~2024

Charlie Giattino, Edouard Mathieu, Veronika Samborska and Max Roser (2023) - "Artificial Intelligence" Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/artificial-intelligence' [Online Resource]

The Journey to AI Pervasiveness in Supercycles



Cloud

Large-Scale Training and Inference

Endpoints

Personal Assistants, Games, Copilot

Embedded

Automotive, Sensor Intelligence, Robotics, Medical

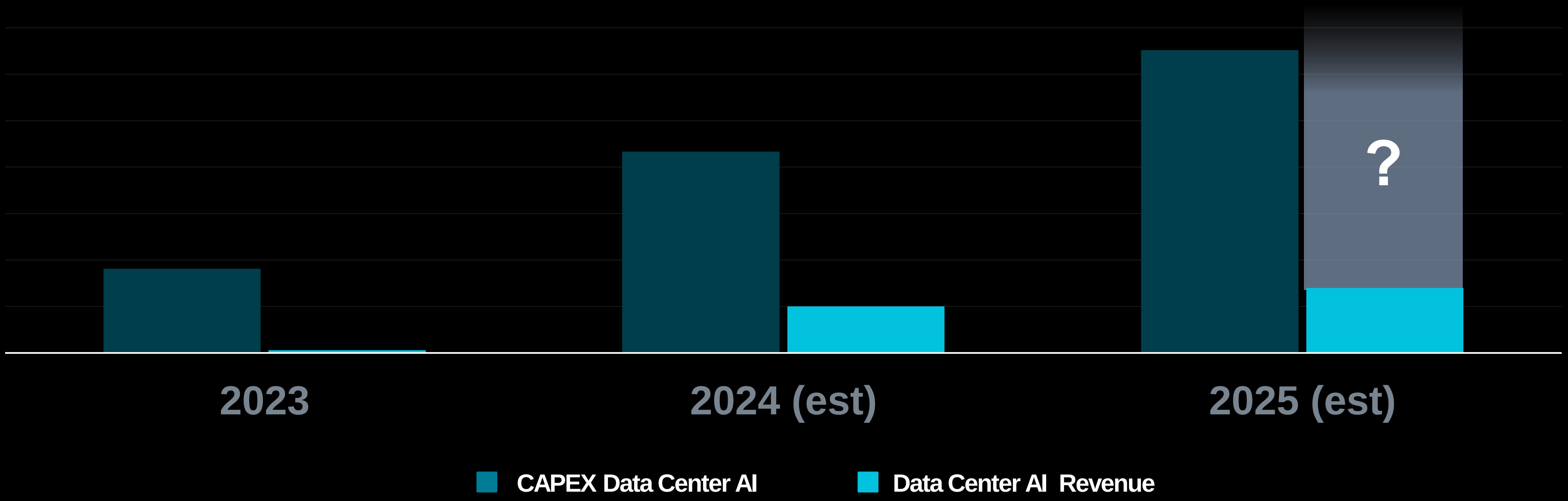
Today's AI

Virtual ❯ Personalized ❯ Physical

With Increasing Unit Volume

AI Pervasiveness

The Investment and Value Creation Race



Today's AI
AI Pervasiveness

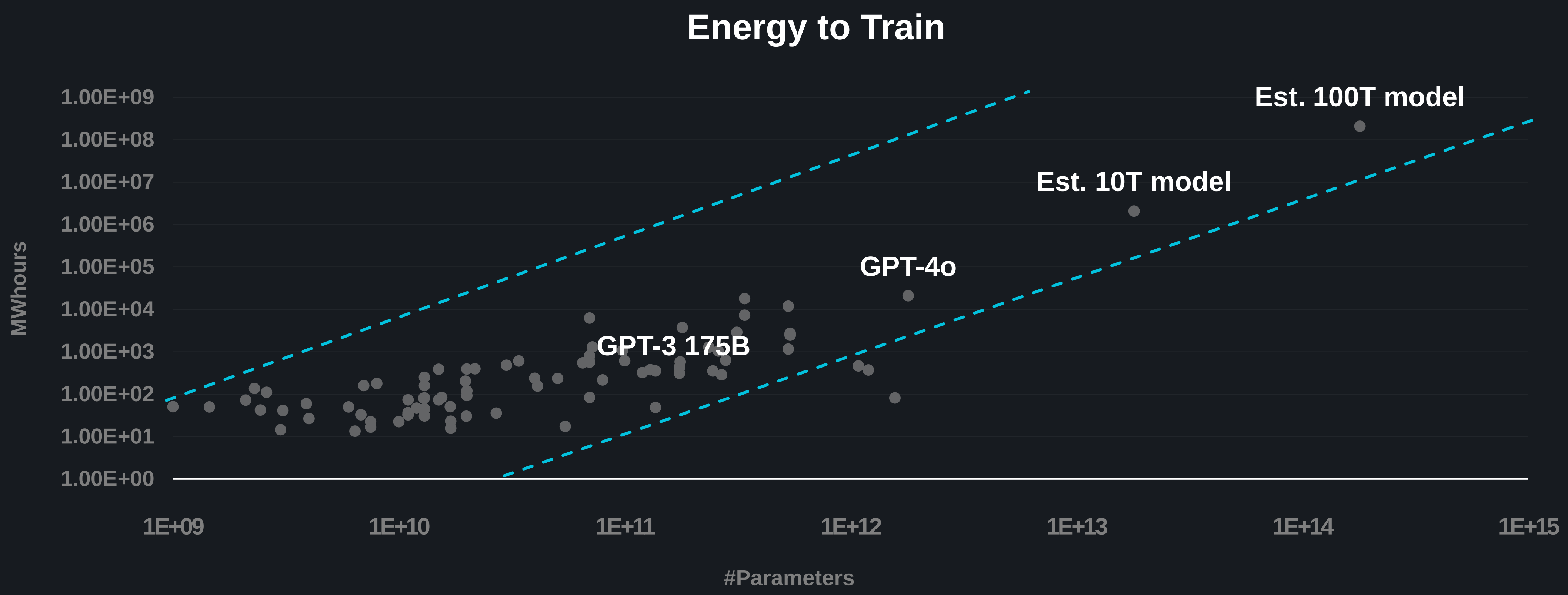
Reported revenue for 2023, internally estimated numbers based on https://www.sequoiacap.com/article/ais-600b-question/ for visualization

The Challenges Ahead

Additional challenges: High quality data, safety, reliability, confidentiality…

On the technology side: Foremost power source & distribution, and diverse requirements of diverse use cases

Power to Train Frontier Models



Exponential growth in model sizes drives massive increase in energy required for training

Total data center power capacity limits model size

Based on published parameter counts of leading training models; and AMD internal calculations approximating energy to train

Energy Efficiency Concerns all Segments

Cloud

- Total data center power capacity limits model size
- Network power contribution is significant (~20%)

Endpoints

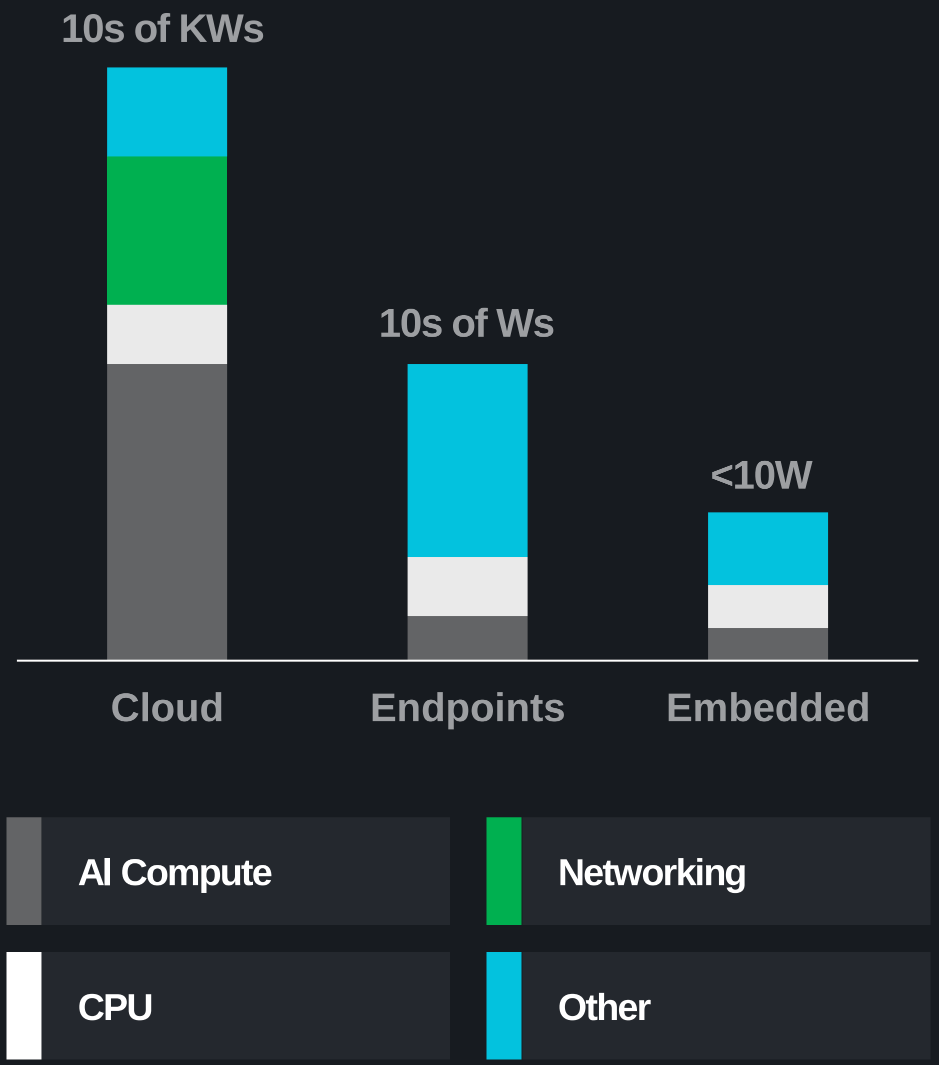
- Trade-off between user experience and battery life

Embedded

- Total power envelop highly constrained

At larger scale we can amortize system overheads for more efficiency

System-Level Breakdown of Power



Based on AMD internal calculations

Different Requirements Across Segments

Cloud AI

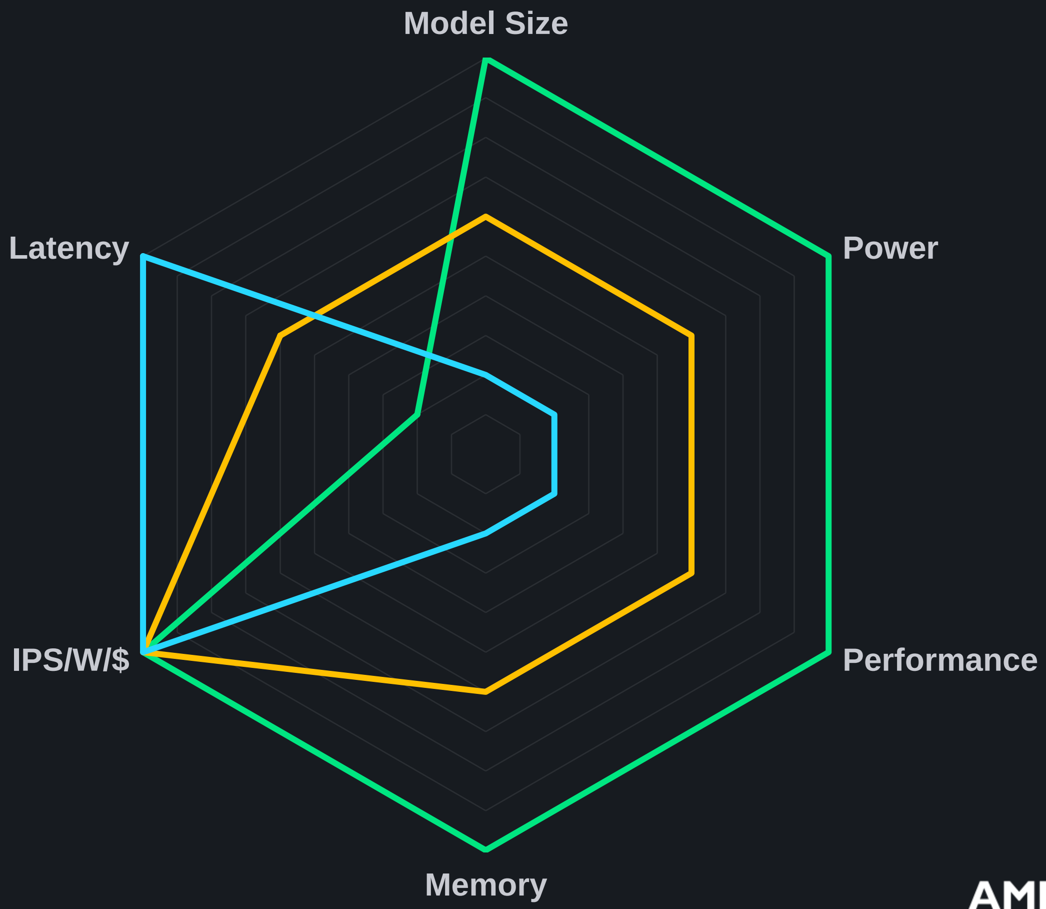
- 1000+ of AI TFLOPS per GPU
- TCO
- Interconnect to thousands of nodes
- Feature velocity critical

Endpoint AI

- 100s of AI TOPS per APU
- Interactive latency
- Highly optimized vendor libraries
- Feature velocity for developers

Embedded AI

- 10s of AI TOPS per SOC
- Real-time requirements driving low latency
- Form factor & functional safety focused workloads
- Customization, I/O



AMD Solutions to Address Key Challenges

1

Energy-Efficient Performance

2

Diverse Requirements

Today’s AI

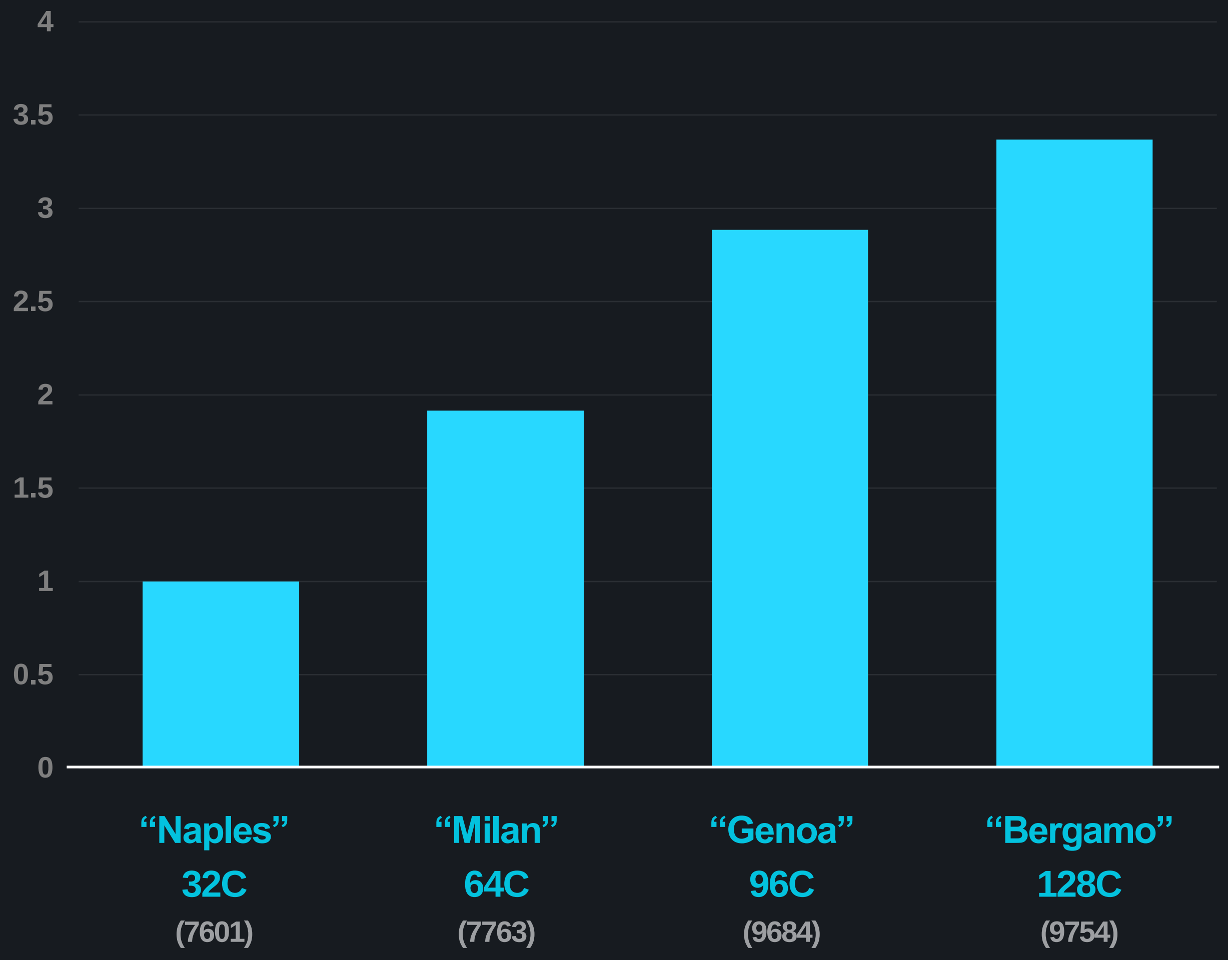
AI Pervasiveness

AMD - Leadership Track Record in Energy Efficiency

A decade of focus on energy efficiency

Leadership efficiency products in server, client, graphics and HPC

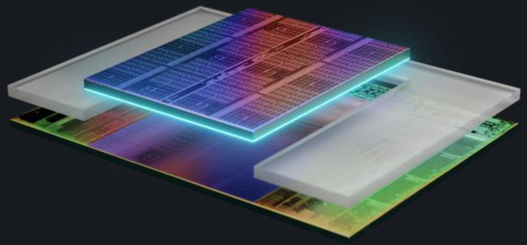
Performance/Watt



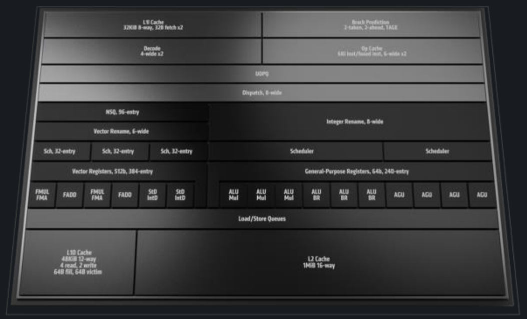
*Normalized to Naples 32C (7601), based on EPYC processor TDP and SpecIntRate2017
Naples: 141, 180W TDP; Milan: 424, 280W, Genoa: 904, 400W, Bergamo 948, 360W
https://www.spec.org/

AMD
together we advance_

Relentless Focus on Performance and Energy Efficiency – All Levels

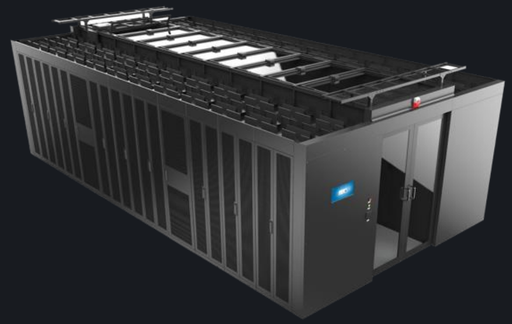


Technology



Architecture

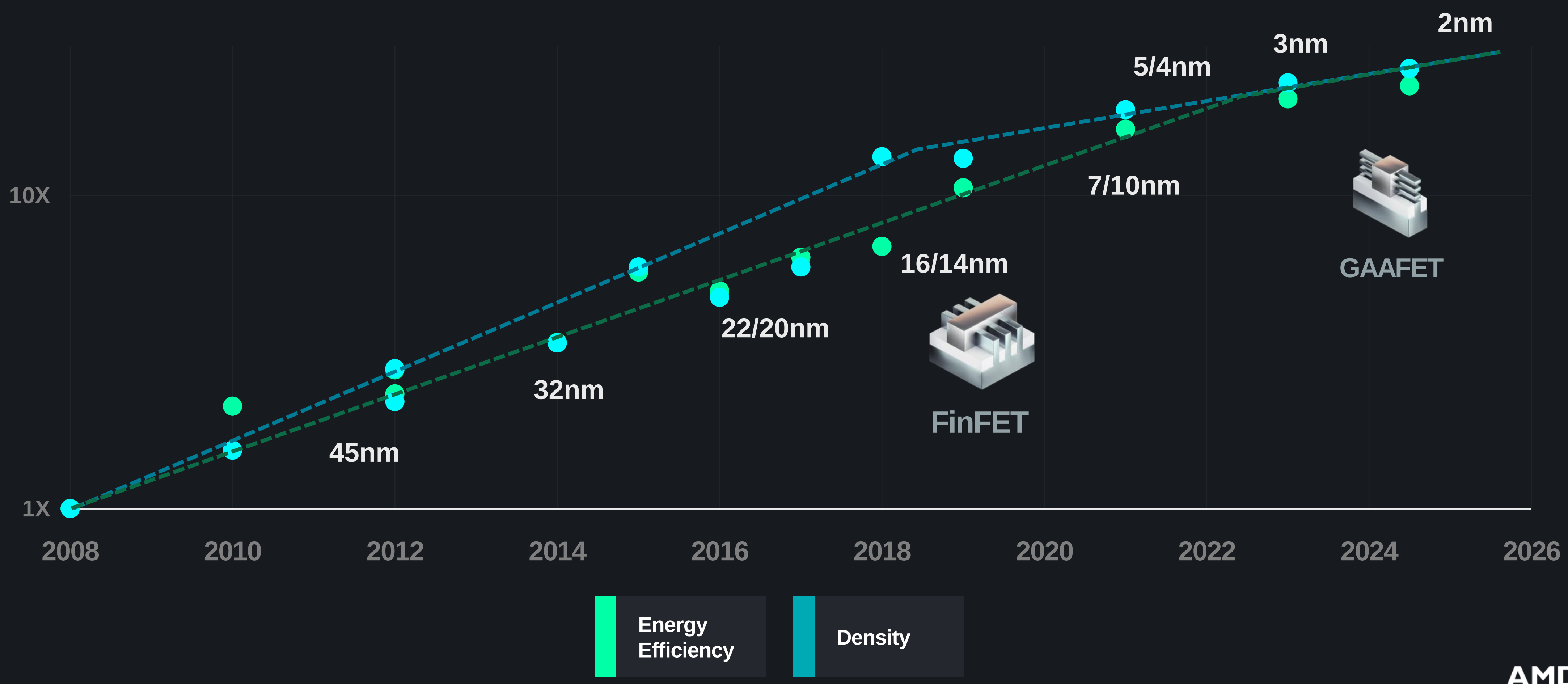
System



DC Infrastructure

Energy Efficiency Gains at Silicon Level

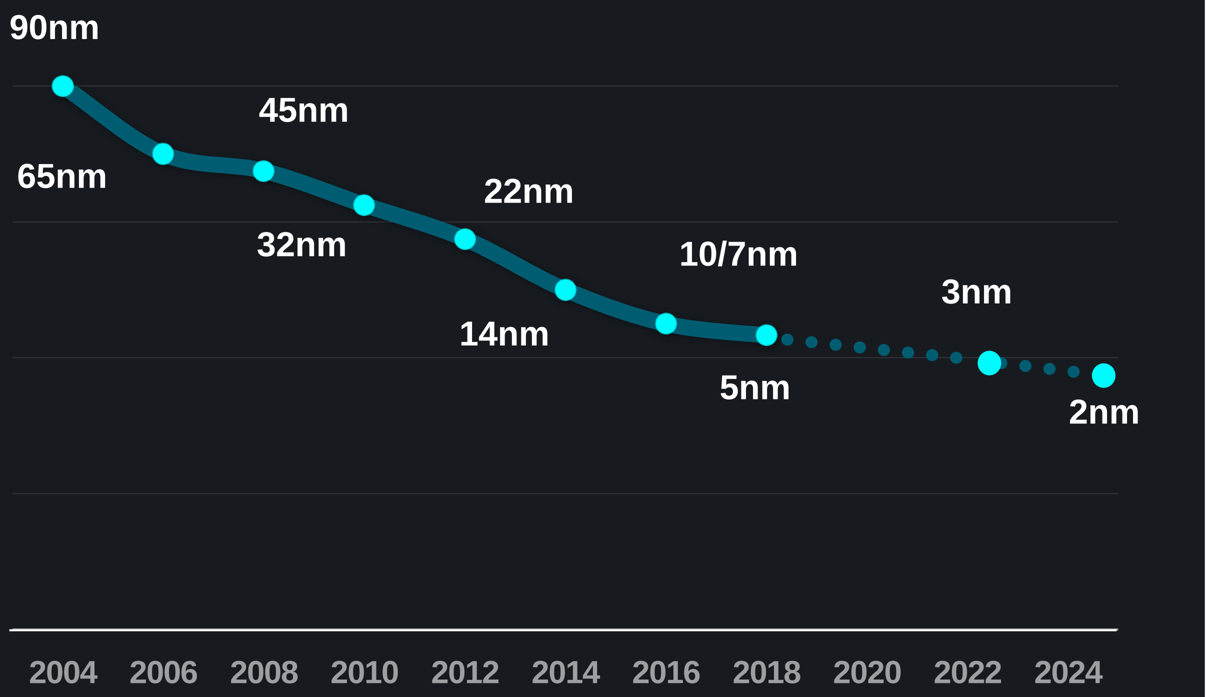
Technology gains slowing but essential



Based on AMD internal estimates

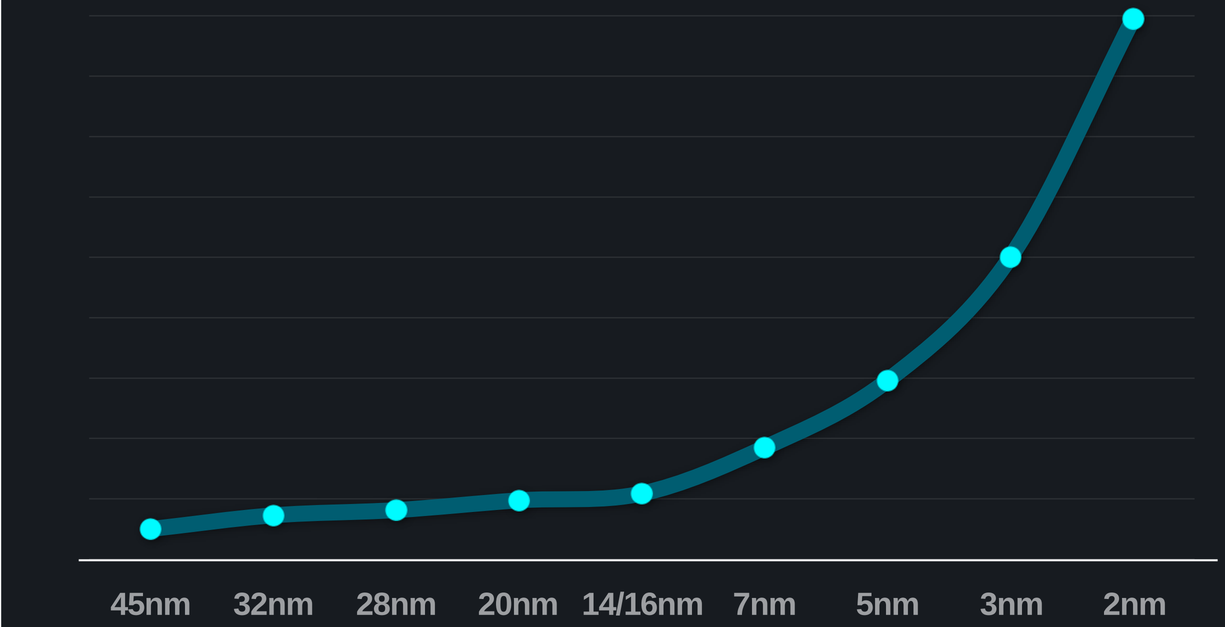
Cost Scaling Challenges

New node introduction rate is slowing



While costs continue to increase

(Cost per yielded mm^2 for a 250mm^2 die)



Source: Internal AMD Data

Advanced Packaging: Significant Gains in Energy-Efficient Performance

Energy-efficient performance needs tight integration

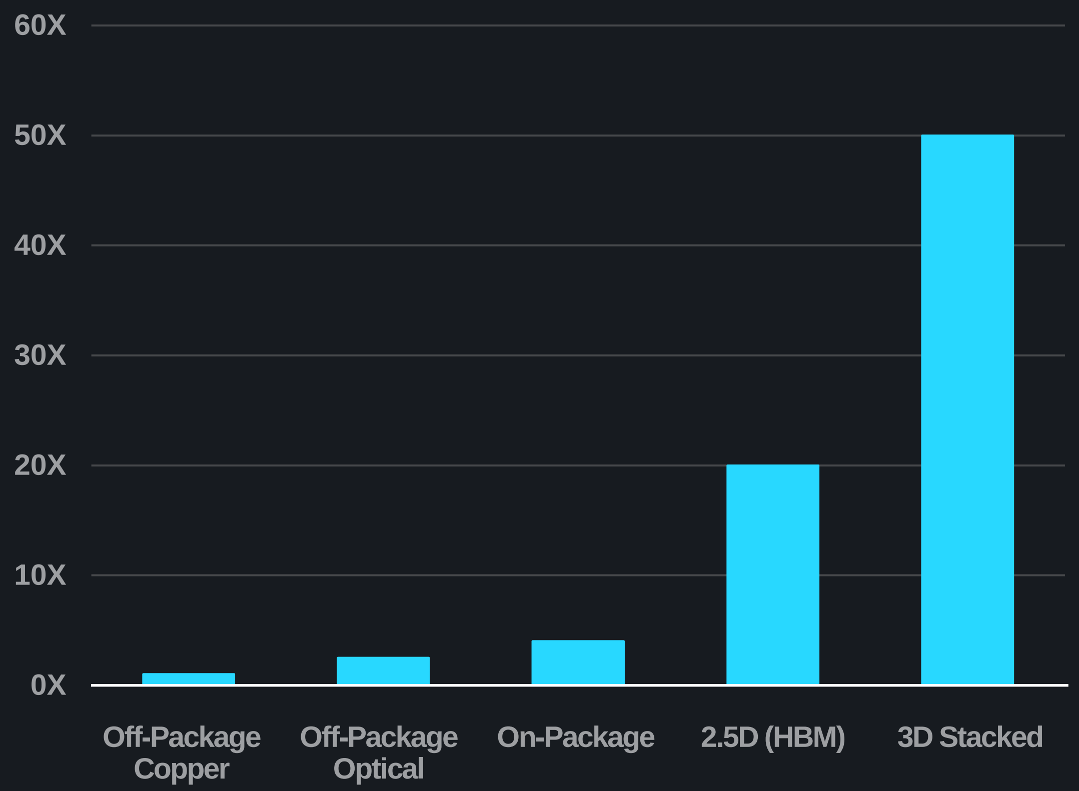
2.5D enables co-packed compute with HBM

AMD 3D V-Cache™ technology drives energy efficiency leadership

Advanced 3D hybrid bonding provides by orders of magnitude the densest, most power efficient chiplet interconnect through higher bandwidth and lower latency

L. Su and S. Naffziger, "1.1 Innovation For the Next Decade of Compute Efficiency," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 8-12, doi: 10.1109/ISSCC42615.2023.10067810

Relative Bits/Joule



Source: Internal AMD Data

Enabling Further Efficiency through Advanced Quantization

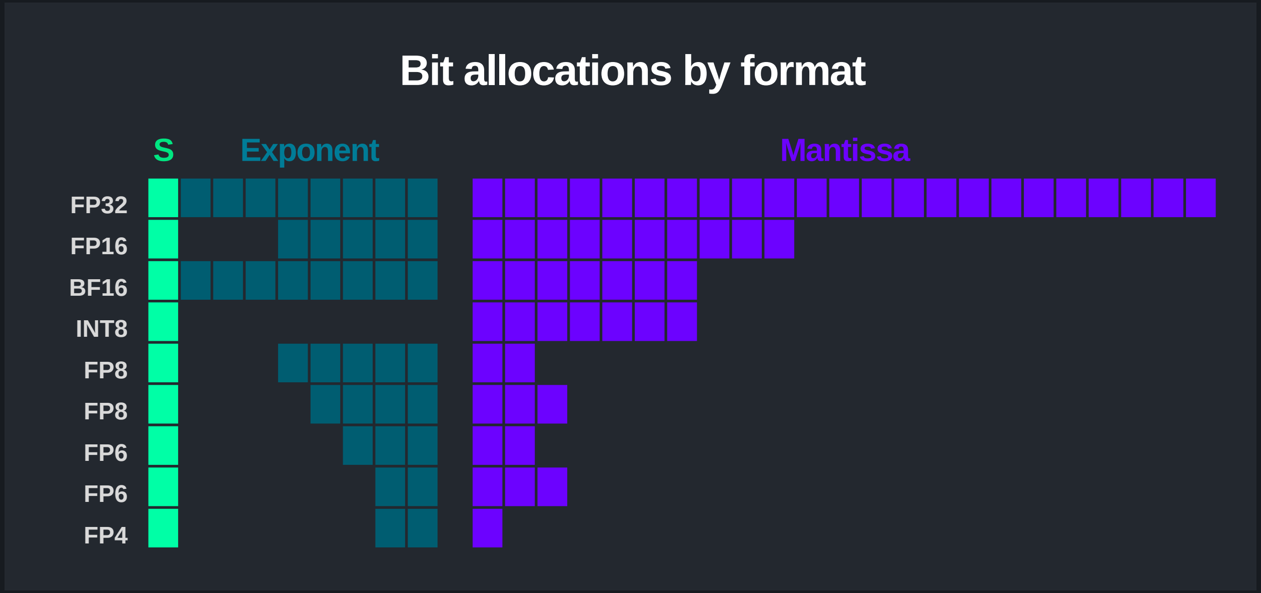
Adapt algorithms to use lower precision math for significant improvements in energy efficiency

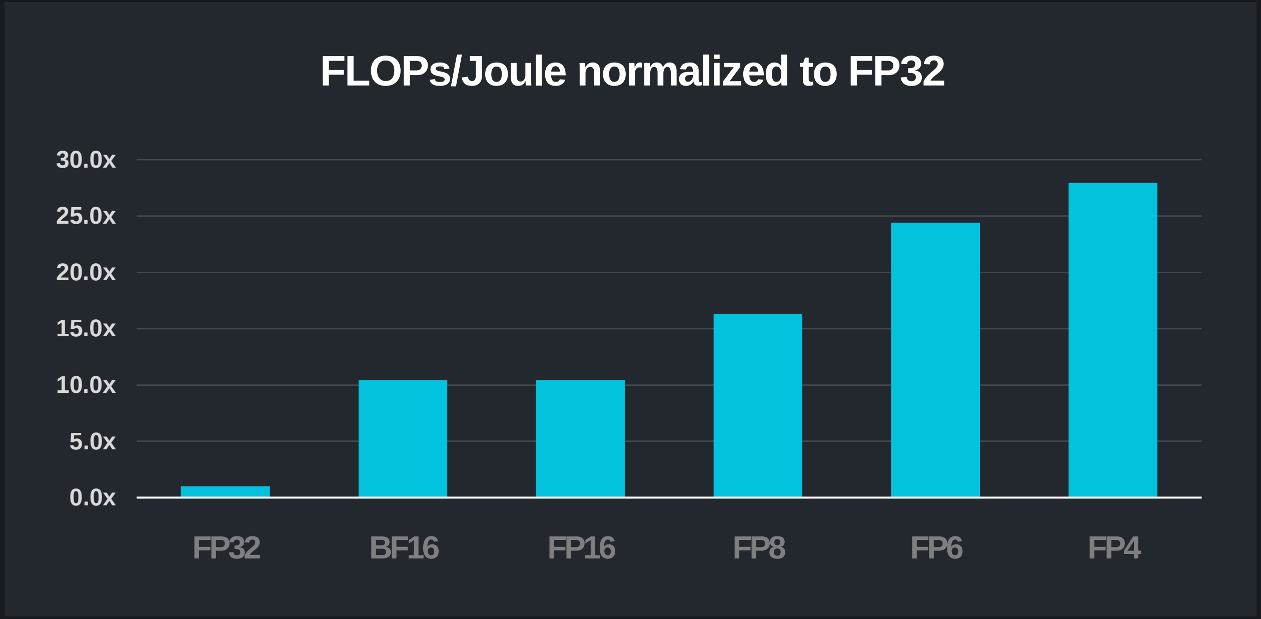
- FP8 in MI325X and FP4 and FP6 in MI350X

Advancing innovation in quantization

Novel research into accumulator-aware quantization[1]

Collaborating through open-source[2]





Source: Internal AMD Data

AMD together we advance_

[1] https://arxiv.org/abs/2301.13376
[2] *brevitas/src/brevitas_examples/imagenet_classification/a2q

Efficiency through Specialization with Dataflow

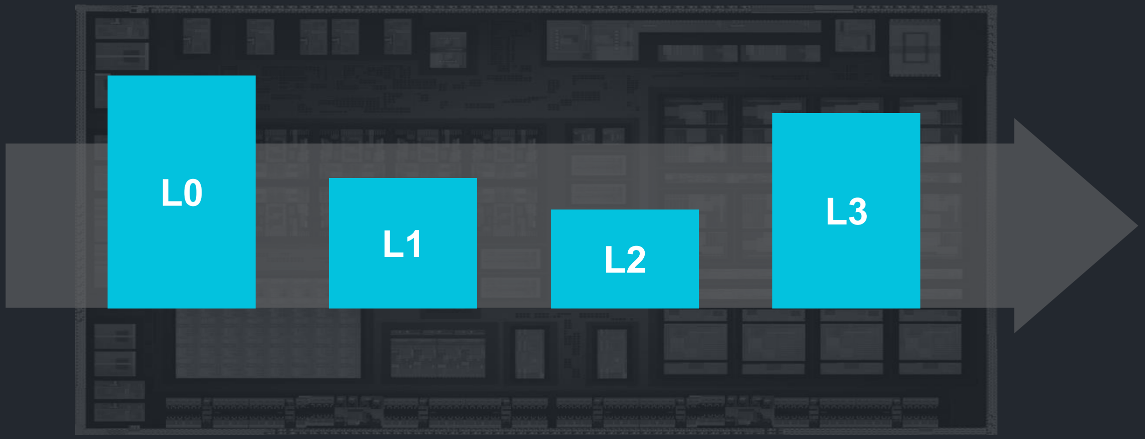
AI model mapped as dataflow architecture

Eliminates intermediate buffering between layers

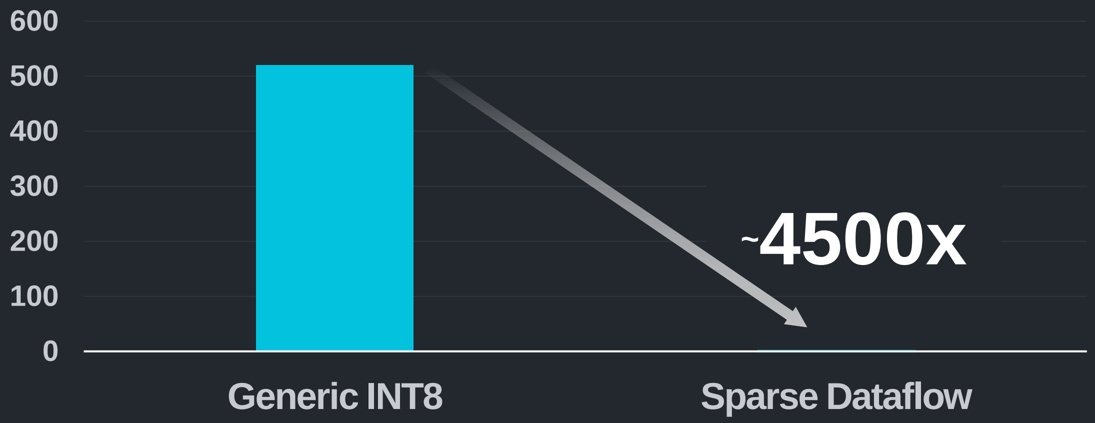
Additional opportunities/research

- Fine-granular customization of datatypes, operations and connectivity
- Potentially replicating irregular sparse compute graphs
- Demonstrates huge potential with ~4500x on FPGAs

DNN Dataflow Execution Architecture



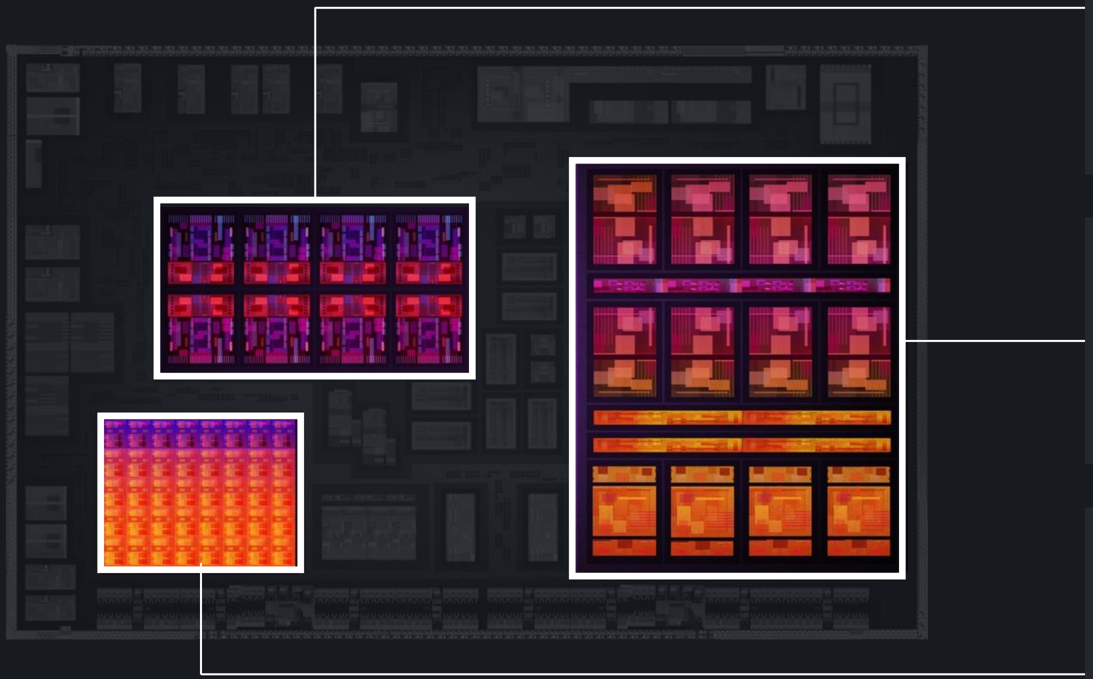
Energy per Inference [uJoules]



Energy calculated as inference time * power consumption ZCU104
Leverages LogicNets and assumes equivalent power to generic implementation
M.Blott “Pervasive and Sustainable AI with Adaptive Computing Architectures” DVCON’2023

Efficiency through Heterogeneity

3rd Generation AMD Ryzen™ AI Technology



Next-Gen GPU

Up to 16 Compute Units

AMD RDNA 3.5

Next-Gen CPU

Up to 12 Cores, 24 Threads

ZEN 5

Next-Gen NPU

Industry-leading 50 NPU TOPS

AMD XDNA 2

*See endnotes: STX-04

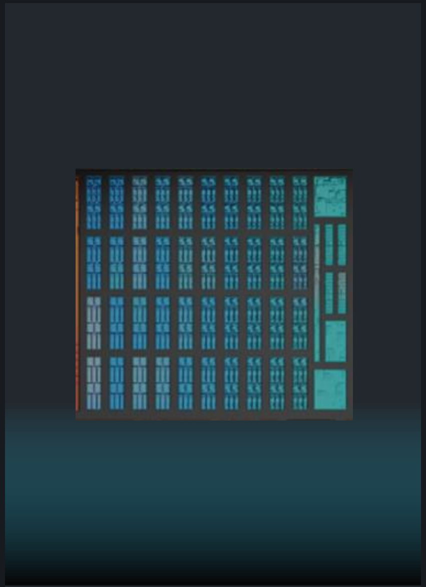
AMD together we advance_

Efficiency in Data Centers: Reducing Data Movement Energy

Networking is significant power contributor at data center scale

Maximizing locality at data center level becomes critical to efficiency

Hierarchical interconnects with scale-up and scale-out



1.0

Register file

1600

Scale-up interconnect

5000

Scale-out interconnect

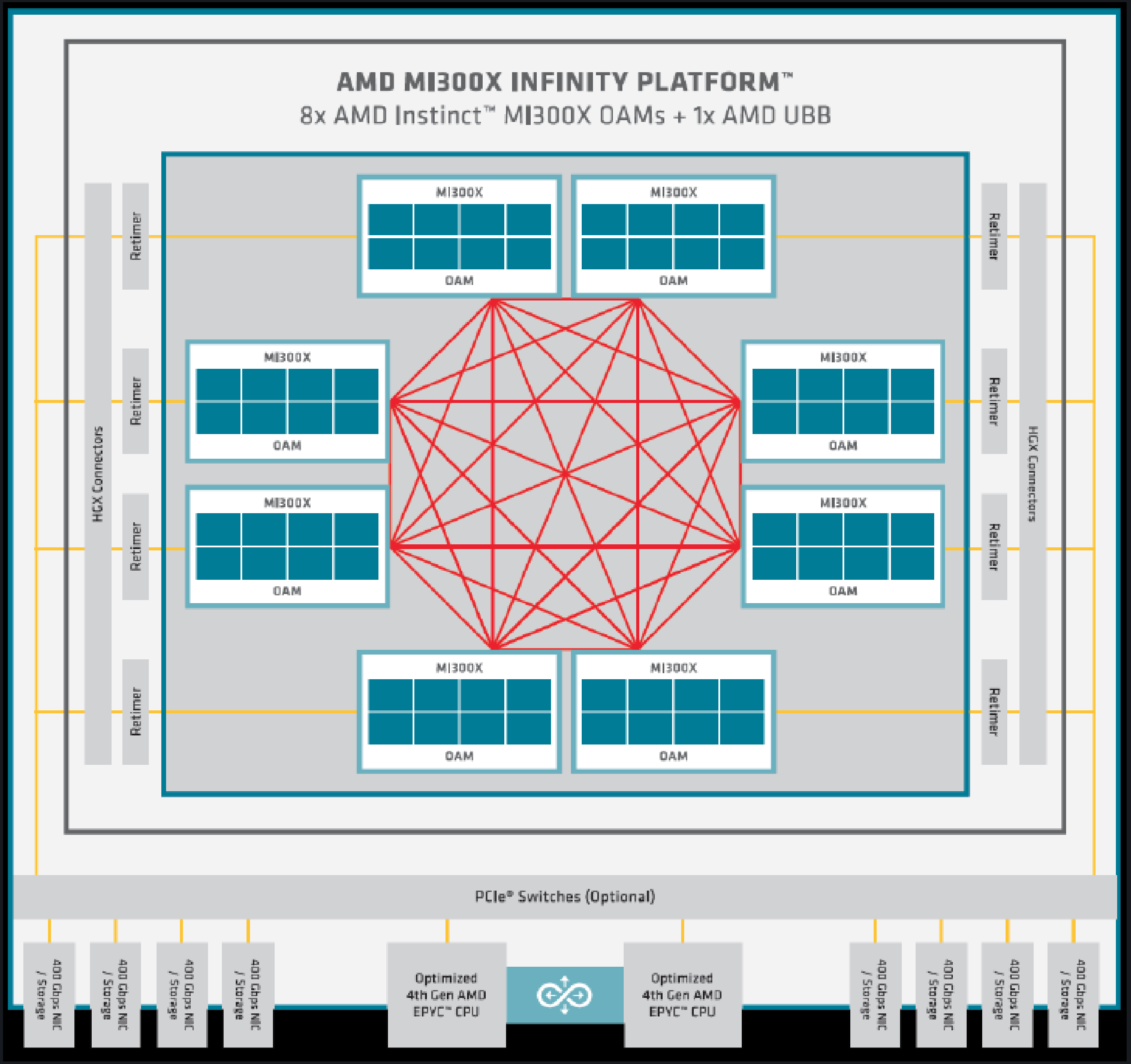
Normalized energy/bit

Source: Internal AMD Data

Efficiency Through Scale-up Interconnects
MI300X Infinity Platform

Direct connectivity for 8 OAMs via AMD Infinity Fabric™

896 GB/s AMD Infinity Fabric™ Bandwidth



AMD MI300X INFINITY PLATFORM™
8x AMD Instinct™ MI300X OAMs + 1x AMD UBB
MI300X
OAM
HGX Connectors
Retimer
PCIe® Switches (Optional)
400 Gbps NIC / Storage
Optimized 4th Gen AMD EPYC™ CPU

AMD
together we advance_

AMD Solutions to Address Key Challenges

1

Energy-Efficient Performance

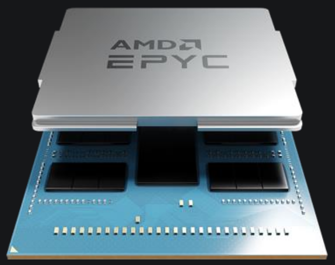
2

Diverse Requirements

Today’s AI

AI Pervasiveness

Broad Portfolio to Address Diverse Spectrum of Requirements

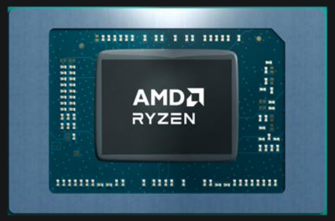


4th Gen
AMD EPYC™
Processors

AMD Instinct™
AI Accelerators

AMD Versal™
FPGAs and SoCs

AMD Radeon™
GPUs

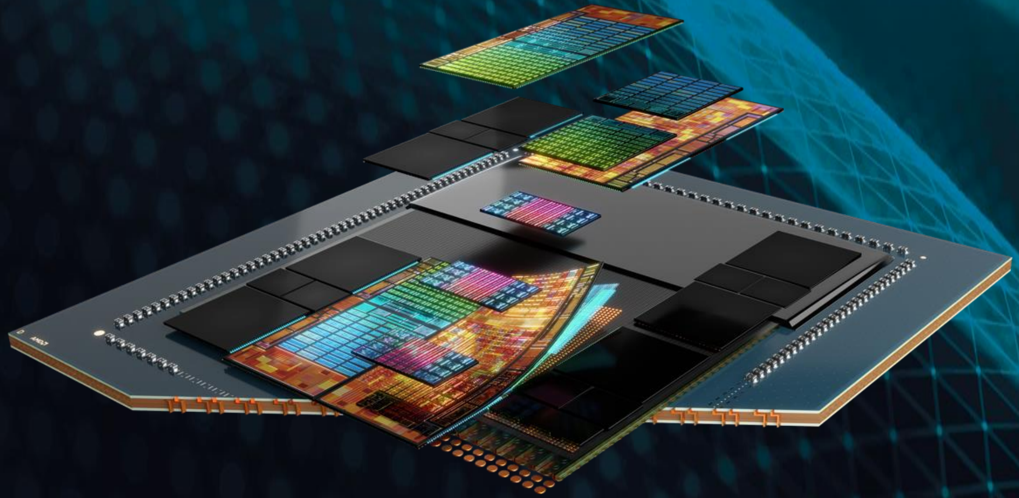


AMD Ryzen™
Mobile Processors

From Cloud to Embedded

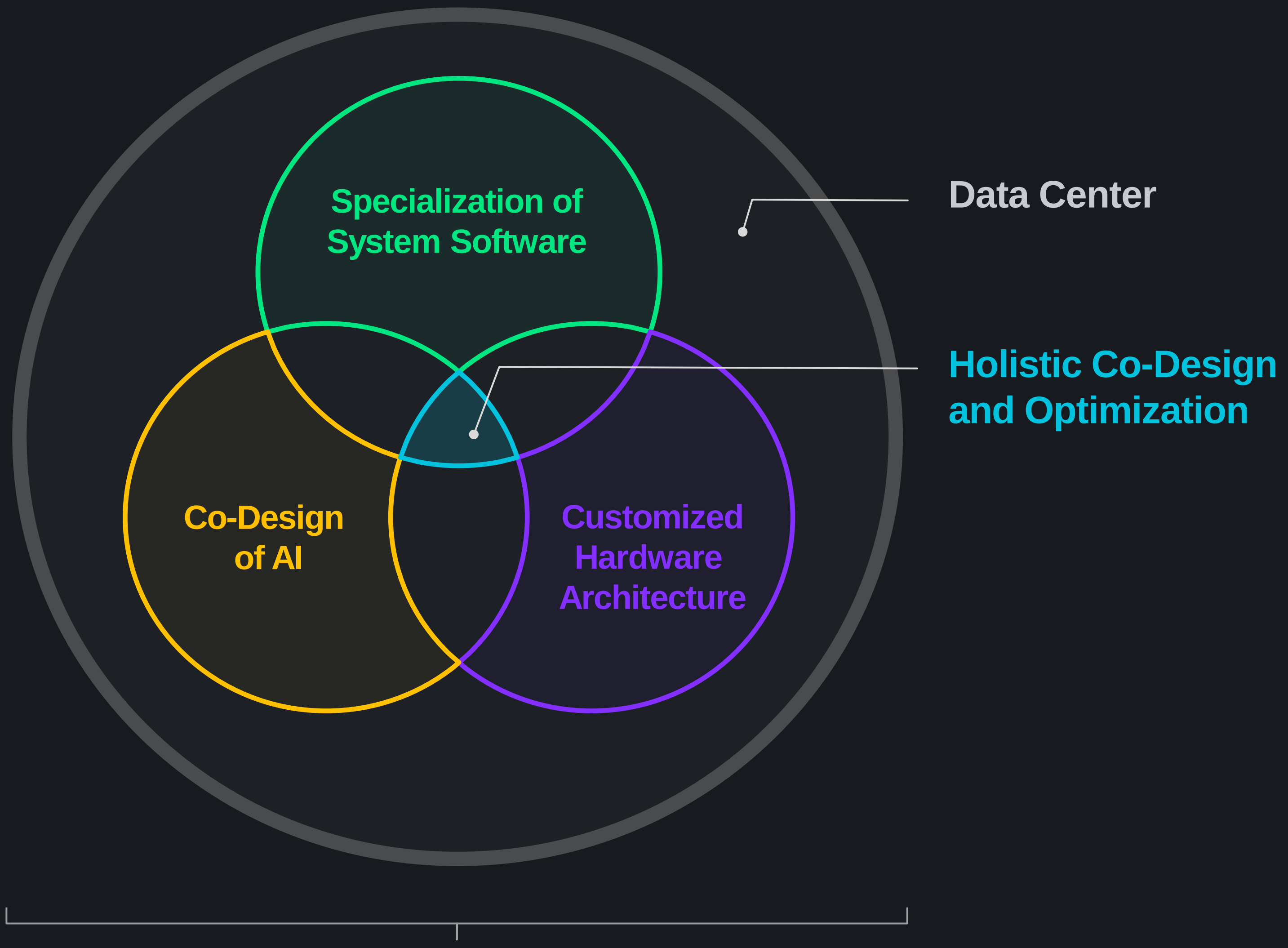
Device Diversity with Leadership in Chiplet and Packaging Technology

- Long history in modularity through chiplet and advanced packaging
- Leveraging technology across markets
- Accelerated enabling of new device families to cater for emerging diverse AI requirements



AMD
together we advance_

Holistic Co-Design and Optimization to Unlock Full Efficiency and Performance


Specialization of System Software
Data Center
Holistic Co-Design and Optimization
Co-Design of AI
Customized Hardware Architecture

From model parallelization, training libraries, compiler stacks, kernel libraries, runtimes, to hardware architectures

AMD
together we advance_

Open Ecosystem Driving Accelerated Innovation

Open-source research, code sharing, standards and datasets enabled critical innovation in AI

- ImageNet
- PyTorch
- “Attention is all you need”
- Llama3

Committed to open collaboration and community innovation

Hugging Face

700,000+ models run out-of-box on AMD ROCm™ platform

PyTorch

Fully upstreamed AMD ROCm™ platform support

Quantization Libraries such as Brevitas

OpenAI Triton

Fully upstreamed AMD ROCm™ platform support

Used for key LLM kernel generation

JAX

vLLM

TensorFlow

MLIR IREE

ONNX Runtime

OpenXLA

AMD
together we advance_

On the Way to AI Pervasiveness
Active Adoption from Cloud to Embedded

Adobe MAXON EA Midjourney BUFFERZONE zoom bitdefender secure your every bit RADiCAL NetEase Games ARKRUNR BORISFX

Sony Interactive Entertainment UNREAL ENGINE SOLIDWORKS blender Camo splashtop Topaz Labs voicemy.ai INTERWOVENWORLDS 游戏科学 GAME SCIENCE

grammarly GoPro LM Studio nero Blackmagicdesign webex by cisco WHISPP Avid convai OBS Studio G+ msi

ADVANTECH ASRock acer ABB AISIN ASUS CERN clarius Continental DELL HESAI hp Hitachi Energy HYPER ACCEL

illumina iBASE innodisk JOHN DEERE JR九州 kakao i cloud Lenovo Microsoft mindray MITSUBISHI ELECTRIC PLANET SPARK PROPHESEE METAVISION FOR MACHINES

NASA NEC RADAR robosense BEYOND SICK Sensor Intelligence. SK telecom SIEMENS SUBARU SUN SINGAPORE tanway Tattile TESLA TAI TOKYO ARTISAN INTELLIGENCE veoneer ZF

Today's AI

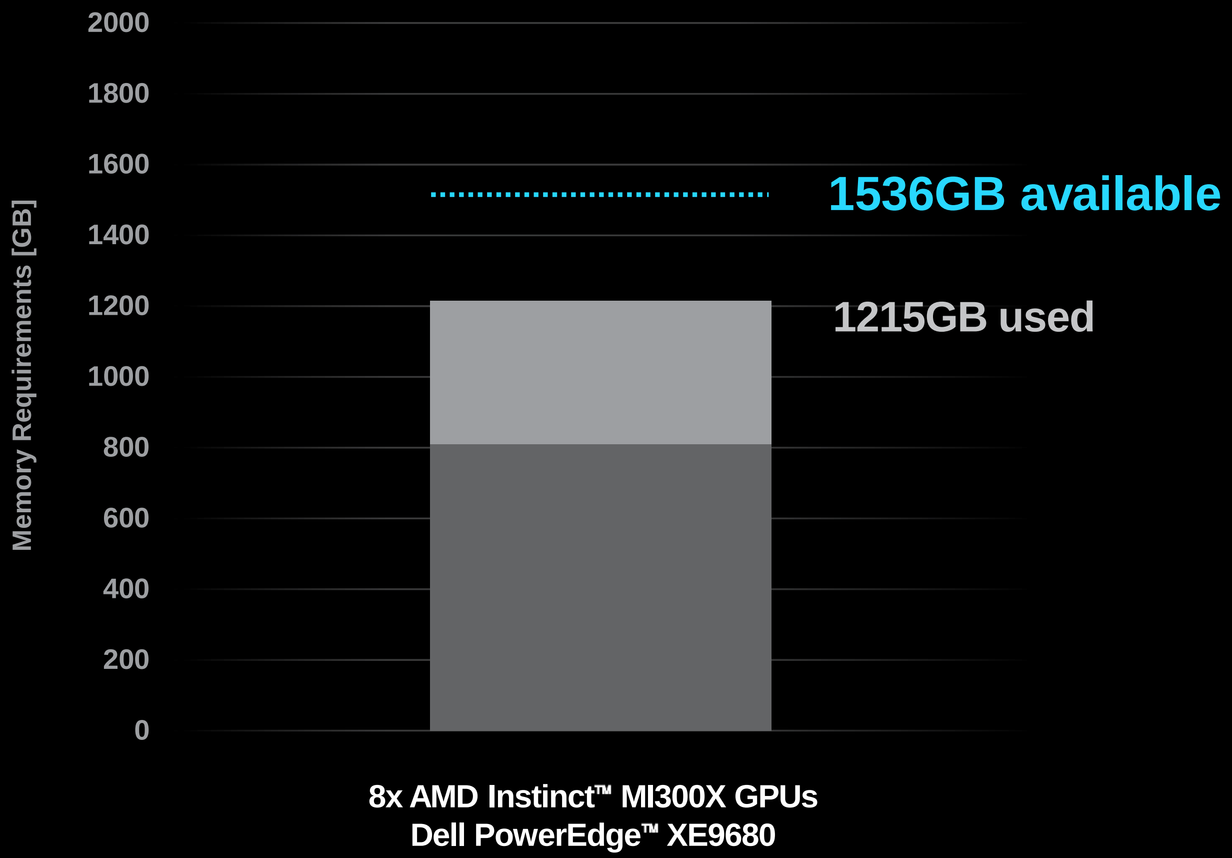
AI Pervasiveness

Meta ORACLE SUPERMICRO GIGABYTE ingrasys Inventec QCT wistron Crusoe Cirrascale cisco BROADCOM ARISTA wiwynn

Aligned DENVR dataworks TENSORWAVE ARKON ENERGY databricks DeepLearning.AI essential AI LAMINI PyTorch

ONNX TensorFlow EVIDEN an atos business

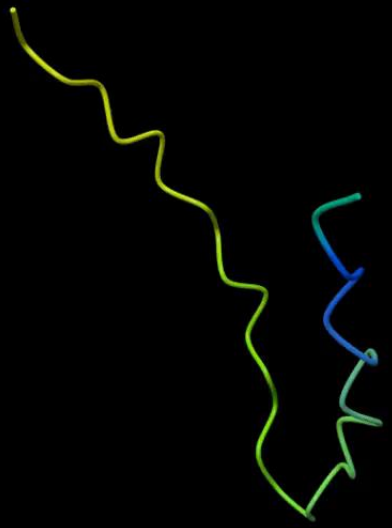
Llama 3.1 enabled by MI300X GPUs



Llama 3

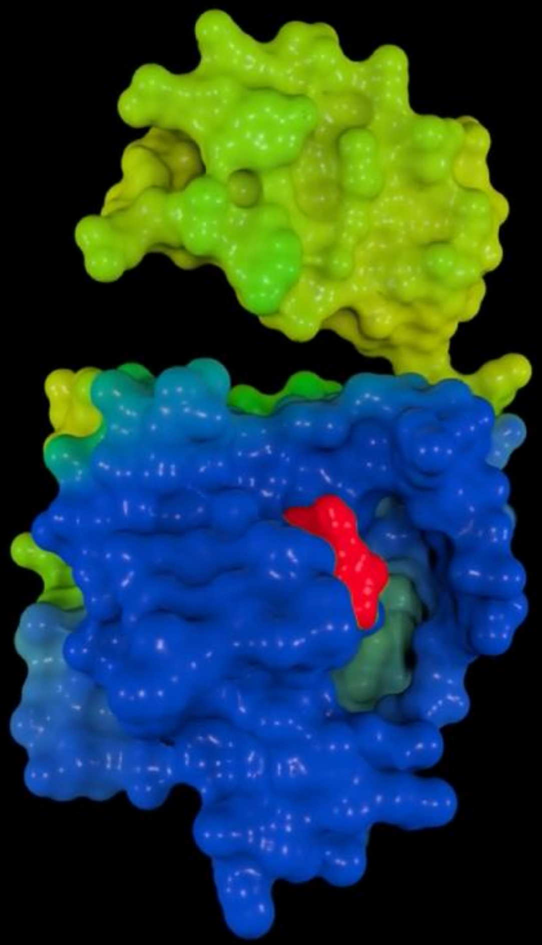
Adopted from: https://infohub.delltechnologies.com/en-us/p/amd-instinct-tm-mi300x-accelerators-on-poweredge-tm-xe9680-rack-server-with-lora-fine-tuning-vllm-model-serving/

- Democratizing Frontier-Level AI Through Open Source
- Energy Efficiency Through Compact Parameter Footprint
- 405B on a single 8-GPU server



Program a better asparaginase protein to starve cancer cells of asparagine

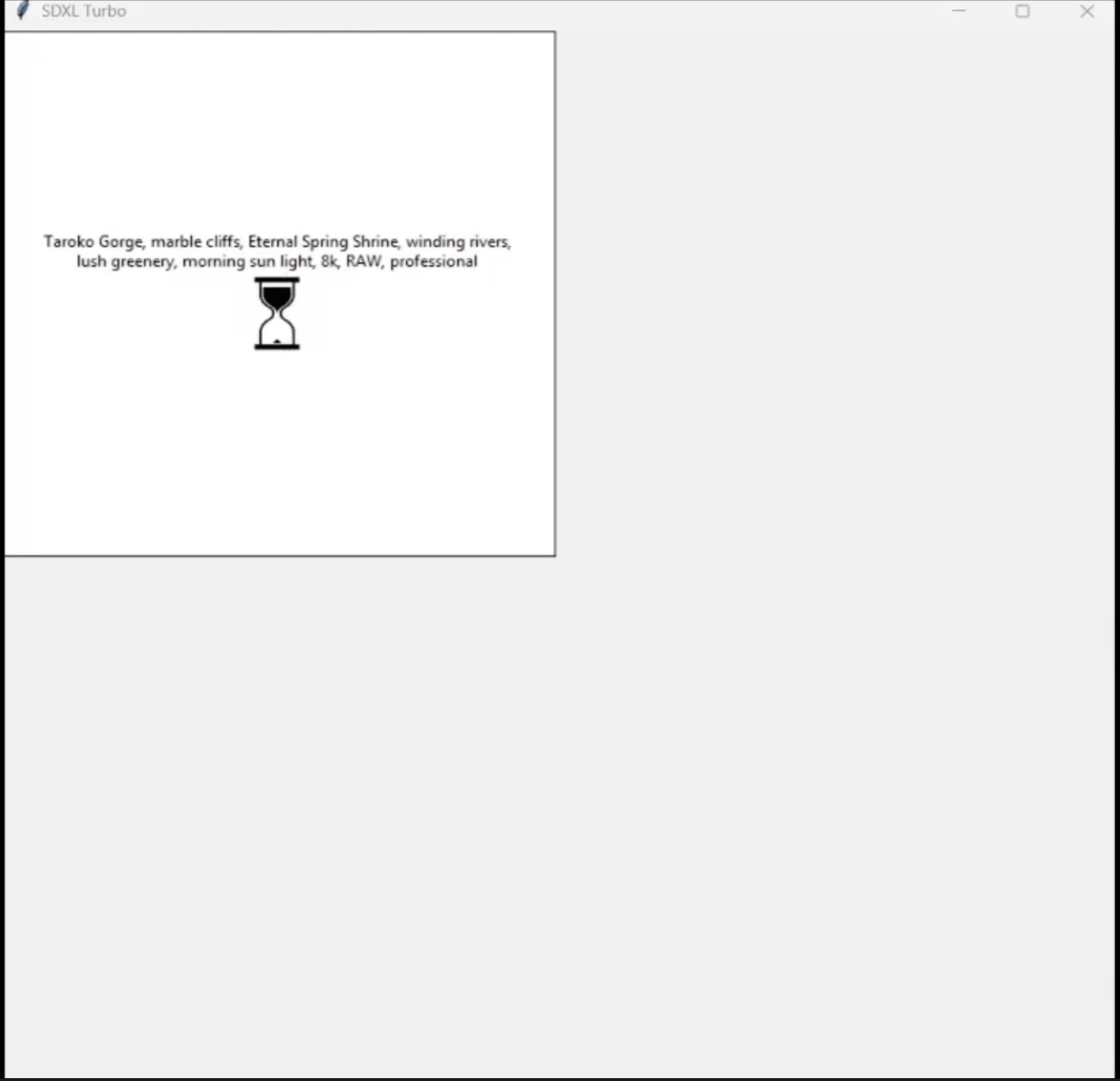
Generative AI for Molecule Programming Accelerated by an AMD EPYC™ processor and AMD Instinct™ GPU based cloud infrastructure



310 Foundation:
AI revolutionizes biomachine Programmability.

Generative AI for Molecule Programming Accelerated by an AMD EPYC™ processor and AMD Instinct™ GPU based cloud infrastructure

AMD
together we advance_



SDXL Turbo

Taroko Gorge, marble cliffs, Eternal Spring Shrine, winding rivers, lush greenery, morning sun light, 8k, RAW, professional

stability.ai

Text-to-image generation excels in photorealism, processing complex prompts, and generating outstanding results

Advanced by AMD Ryzen™ AI technology

AMD
together we advance_

radmantis

Kria™ SOM - FPGA based SOC platforms: Accelerating AI in real-time at low power in high constraint environments

AMD
together we advance_

Search for Elusive Particles at Europe’s Large Hadron Collider

CERN use case: Needle-in-a-haystack search at extreme scale

- ~ 2.4B collisions/sec producing Pb/s of raw data Real-time machine learning need data analysis

10s thousands of FPGAs 100m underground,

- Each inferencing Tb/s of data with nsec latency

With highly customized dataflow AI accelerators in FPGAs at data-center scale

CERN: Thea Aarrestade (CERN, ETH Zurich): https://nssmic.ieee.org/2024/program/
https://xilinx.github.io/finn/

Extreme throughput and nanosecond latency with highly customized dataflow AI accelerators in FPGAs at data center scale.

AMD together we advance_

Visual created using Midjourney

AMD is delivering full stack performant and power efficient solutions with partners and the ecosystem to customers from cloud to endpoints and embedded

AMD

together we advance_

Endnotes

- GD-83: Use of third-party marks/logos/products is for informational purposes only and no endorsement of or by AMD is intended or implied.

- GD-220a: Ryzen™ AI is defined as the combination of a dedicated AI engine, AMD Radeon™ graphics engine, and Ryzen processor cores that enable AI capabilities. OEM and ISV enablement is required, and certain AI features may not yet be optimized for Ryzen AI processors. Ryzen AI is compatible with all AMD Ryzen 7040 series processors except the Ryzen 5 7540U and Ryzen 3 7440U. Please check with your system manufacturer for feature availability prior to purchase.

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. 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.

ATTRIBUTION