

第24回PCクラスタシンポジウム

Advancing AI: AMDのCPU・GPU 最新情報

2024年12月5日

日本AMD株式会社
大原久樹 (Hisaki.Ohara@amd.com)

Top500 and Green500 (SC24)

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	El Capitan - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	Aurora - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	Eagle - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
5	HPC6 - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461

Green500

3rd: 69.098 GFlops/Watt
GENCI-CINES, France
(440th HPL)

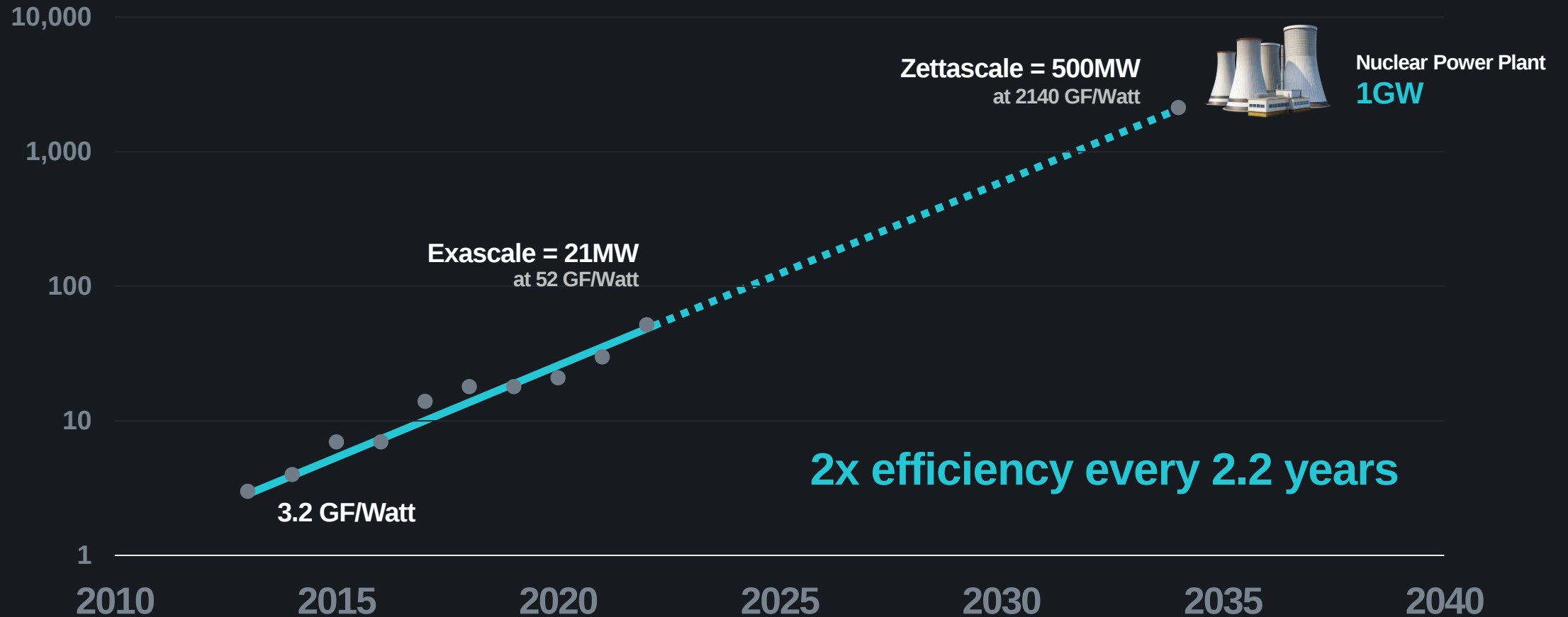
18th : 58.889 GFlops/Watt

22nd : 54.984 GFlops/Watt

21st : 56.483 GFlops/Watt

Supercomputer Energy Use Trajectory

Green500 Supercomputer GFLOPs/Watt and Projection

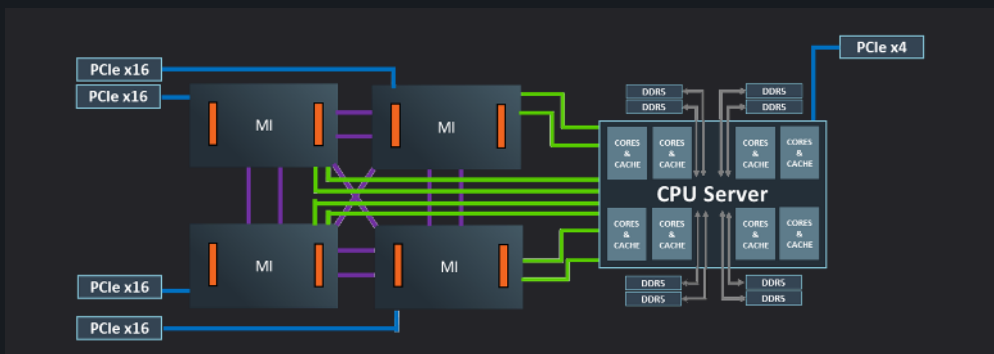


System level energy efficiency optimization

The AMD 30x by 2025 energy efficiency goal - 2024 update

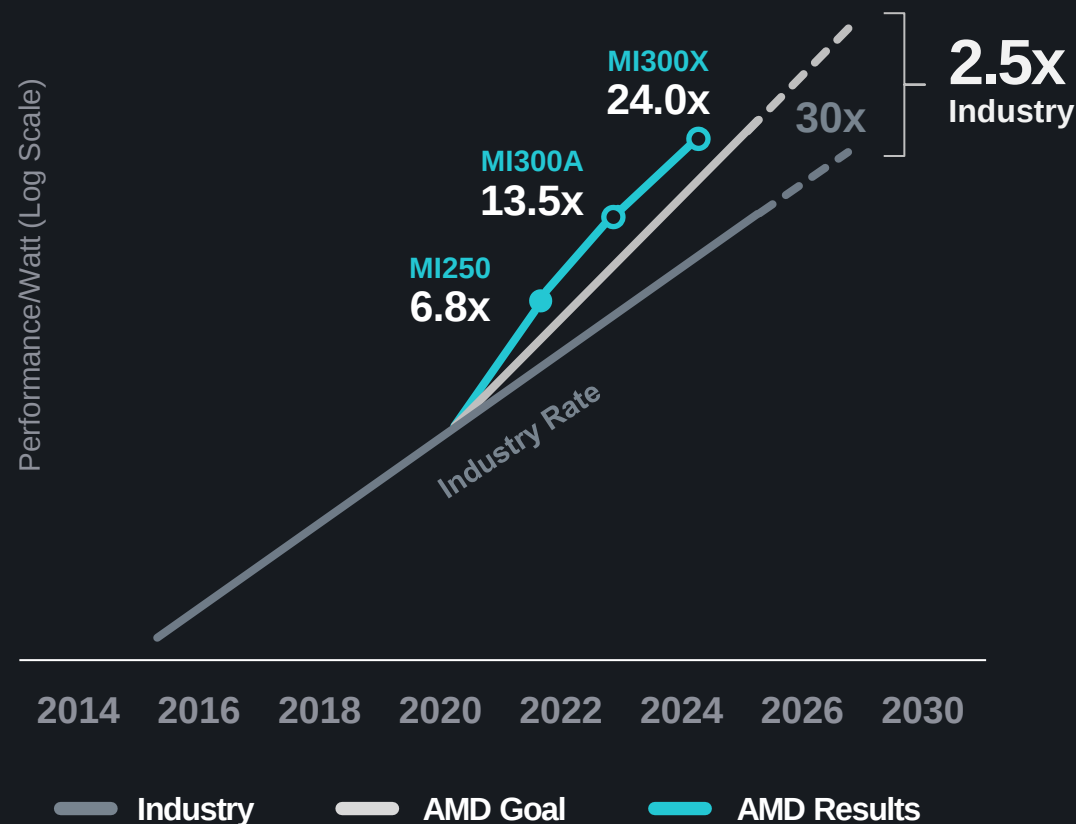
- Goal established in 2021 targeting a 30X increase in efficiency by 2025, much faster than industry trend
- Based on efficiency achieved at the compute node (4 GPU + CPU host) level, rather than device level
- Exploit architectural innovations, package and silicon technology advances to bend the curve to 30x

HPC/ML Node



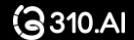
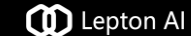
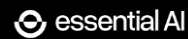
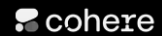
See endnotes

Accelerated computing performance/watt trends



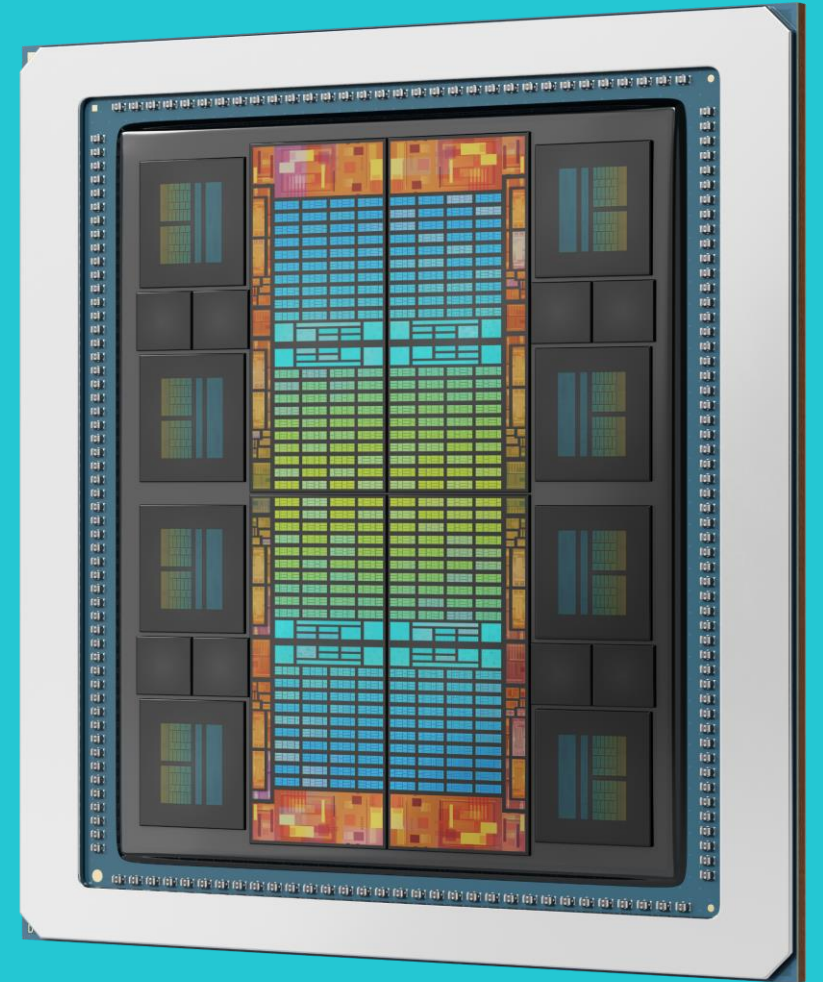
AMD Instinct™ MI300 Series

Powering the most popular Gen AI Solutions



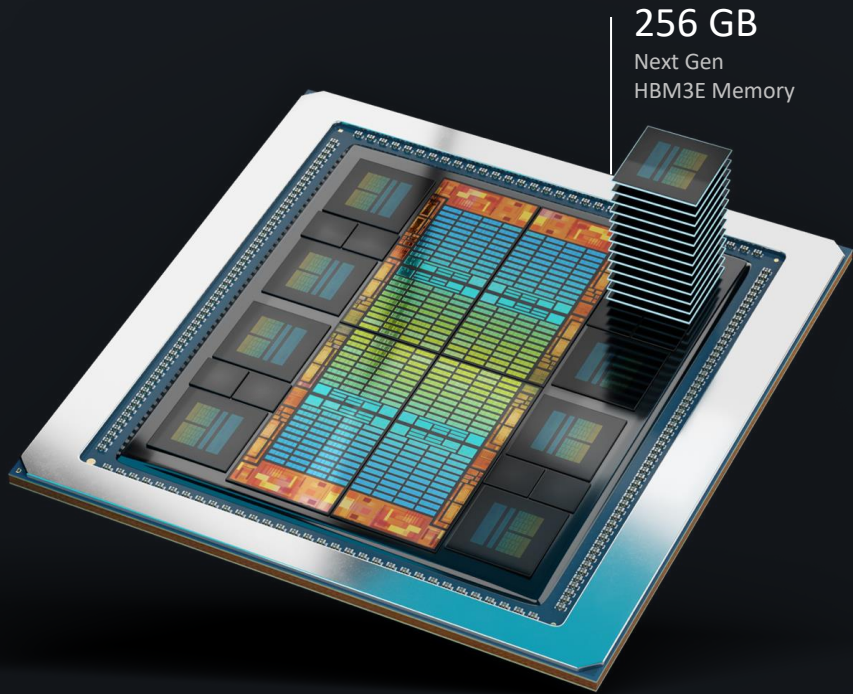


Launching AMD Instinct™ MI325X Accelerators Extending Gen AI leadership



AMD Instinct™ MI325X accelerators

Extending Gen AI leadership



Instinct™ MI325X

AMD CDNA™ 3
GPU Arch.

256GB HBM3E
Memory

2.6 PF
FP8 Perf.

153B
Transistors

6 TB/s
Memory BW

1.3 PF
FP16 Perf.

Note: FP8 and FP16 specs are peak theoretical

AMD
INSTINCT
MI350 Series

Previewing today

AMD Instinct™ MI350 Series

Continued Gen AI Leadership

AMD
CDNA 4

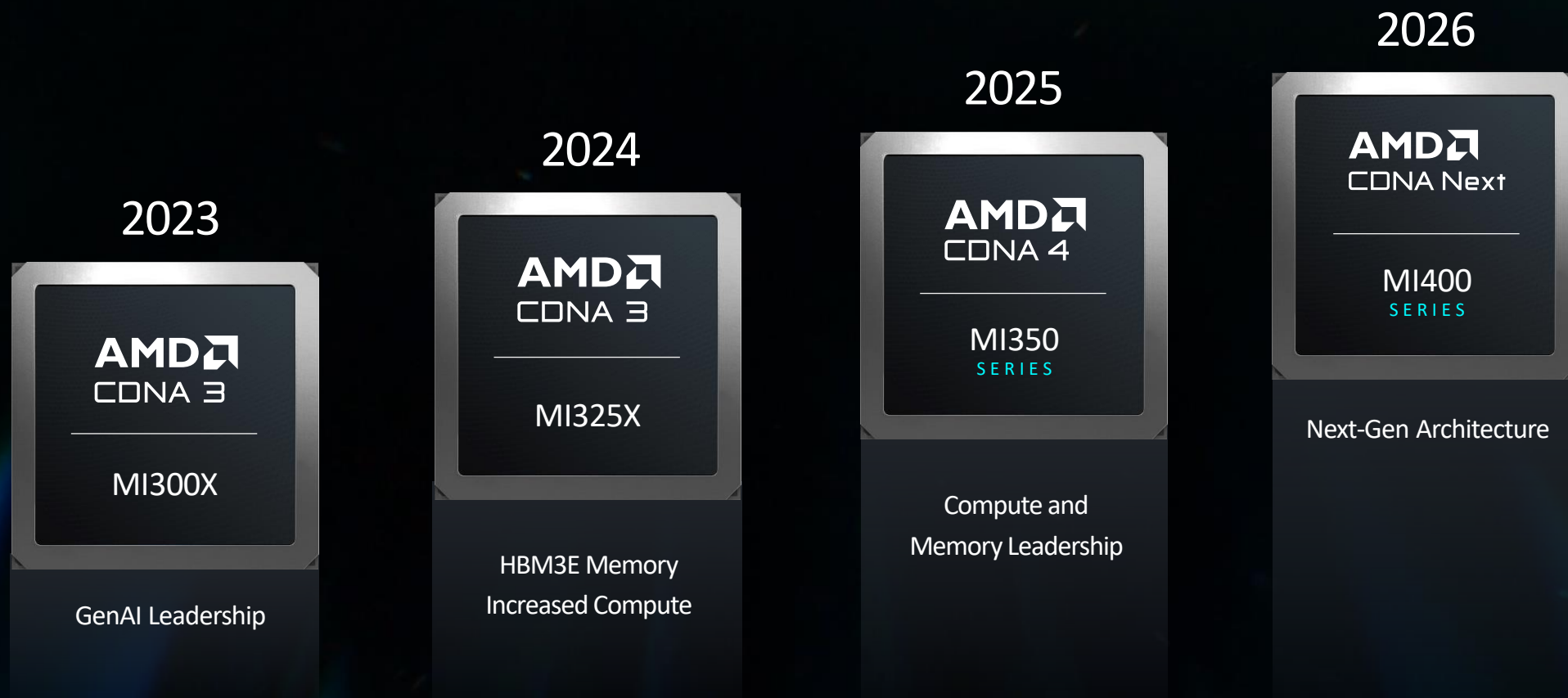
3nm
Process Node

Up to
288GB
HBM3E

NEW
FP4 / FP6
Datatype Support



LEADERSHIP ROADMAP ON AN ANNUAL CADENCE



Fastest growing product in AMD history

AMD Instinct™ MI300X Accelerator

Performant out-of-box support on popular generative AI models

1M+

models supported
out of the box



Hugging Face

Extended support
for leading models

Meta
Llama 3.1

Meta
Llama 3.2



Day 0 support
for AMD GPUs

1.2x

Llama 3 405B
latency improvement

MI300X vs. H100

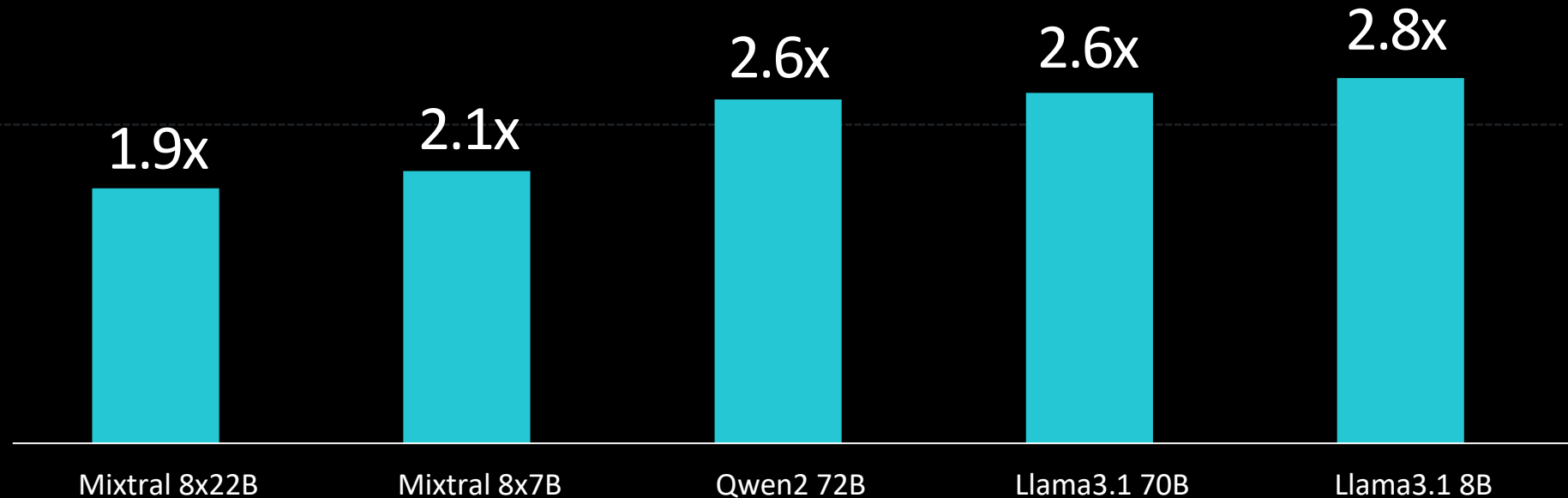
Leadership performance
on popular models

Generational inference improvement

ROCm 6.2 vs. ROCm 6.0

2.4x

average
performance
improvement



Runtime Optimization

Kernel Fusion

Collective Communication

Subgraph

ROCm Fortran Compiler for OpenMP Offloading [preview]

Introducing AMD's Next-Gen Fortran Compiler

2024 November 13 by [Justin Chang](#), Brian Cornille, Michael Klemm, and Johanna Potyka.

We are excited to share a brief preview of AMD's [Next-Gen Fortran Compiler](#), our new open source Fortran compiler supporting OpenMP offloading. AMD's [Next-Gen Fortran Compiler](#) is a downstream flavor of [LLVM Flang](#), optimized for AMD GPUs. Our [Next-Gen Fortran Compiler](#) enables OpenMP offloading and offers a direct interface to ROCm and HIP. In this blog post you will:

1. Learn how to use AMD's [Next-Gen Fortran Compiler](#) to deploy and accelerate your Fortran codes on AMD GPUs using OpenMP offloading.
2. Learn how to use AMD's [Next-Gen Fortran Compiler](#) to interface and invoke HIP and ROCm kernels.
3. See how AMD's [Next-Gen Fortran Compiler](#) OpenMP offloading exhibits competitive performance against native HIP/C++ codes, benchmarking on AMD GPUs.
4. Learn how to access a pre-production build of the new AMD's [Next-Gen Fortran Compiler](#).

<https://rocm.blogs.amd.com/ecosystems-and-partners/fortran-journey/README.html>

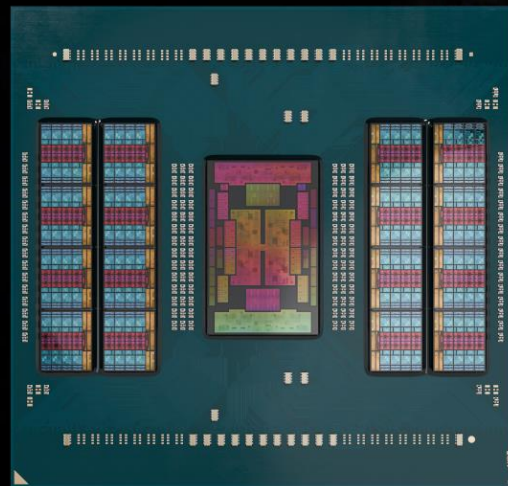
5th Gen AMD EPYC™ Generational Innovations

Compute

- “Zen5” up to **128 cores** / 256 threads
- “Zen5c” up to **192 cores** / 384 threads
- AVX-512 with **full 512b data path**
- New **500W** performance option
- Faster **5GHz** options
- **3/4nm** Zen cores

I/O & Platform

- 2P and 1P Configurations
- Up to 160 lanes of PCIe® Gen5
- **PCIe link encryption**
- SP5 Compatible with “**Genoa**”
- **CXL® 2.0¹**



Memory

- 12 ch. DDR5 ECC **up to 6400* MT/s**
- Up to 2 DIMMs/channel capacity delivering up to **6TB/socket**
- **Dynamic Post Package Repair** (PPR) for x4 and x8 ECC RDIMMs

Security

- Hardware Root-of-Trust
- **Trusted I/O**
- **FIPS 140-3 in process**

1 - CXL Type 1&2 devices and PCIe link encryption support dependent upon ecosystem readiness
*Standard roadmap offerings on AMD.com support 6000 MT/s
As of 10/10/2024. See endnotes 9xx5-048, 072, 083, GD-183A

“Zen 5” Microarchitecture Overview

NextGen Branch Predictor Caches

- I-Cache: 32KB, 8-way; 2x 32B fetch/cycle
- Op-Cache: 6K inst; 2x 6-wide fetch/cycle
- D-Cache: 48KB, 12-way; 4 mem ops/cycle
- L2-Cache: 1MB, 16-way

Dual I-Fetch/decode pipes

- 4 instructions per decode pipe
- 8 ops/cycle dispatch

4 inst/pipe 8 ops/cycle dispatched to Integer or FP Execution capabilities

- 6 integer ALU
- 4 AGU, 4 addresses to LS per cycle
- 6 FP ops/cycle; 2-cycle FADD
- Full 512b AVX512 datapaths

Dataflow

- 4 load pipes capable of 2, 512b AVX512 loads
- 2x width L2 cache <-> L1I and L1D caches

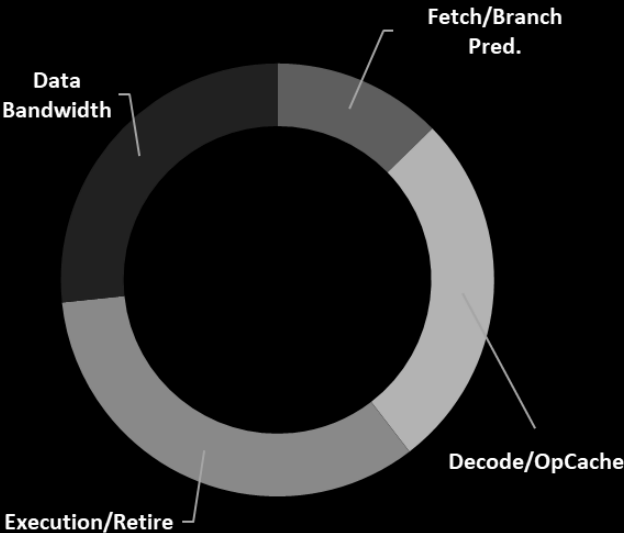
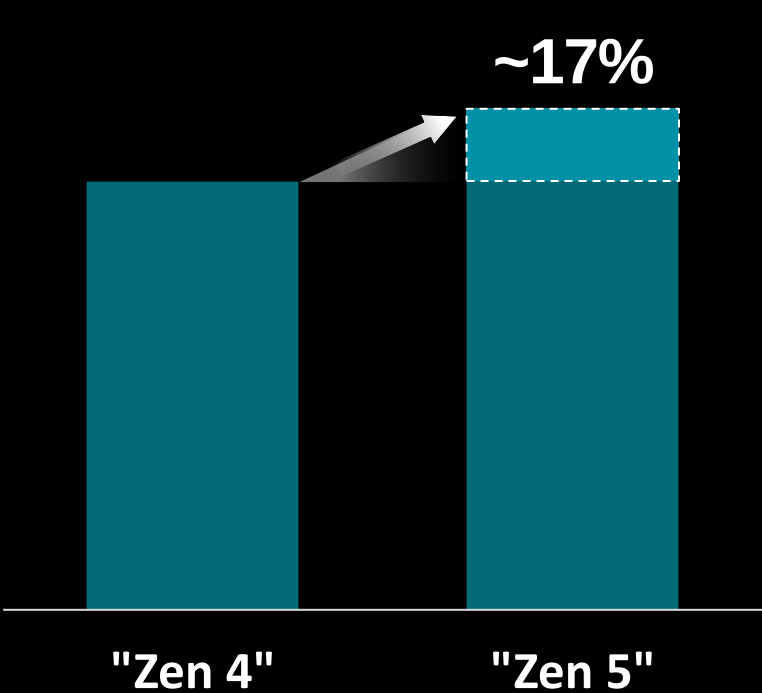
2 Threads per core



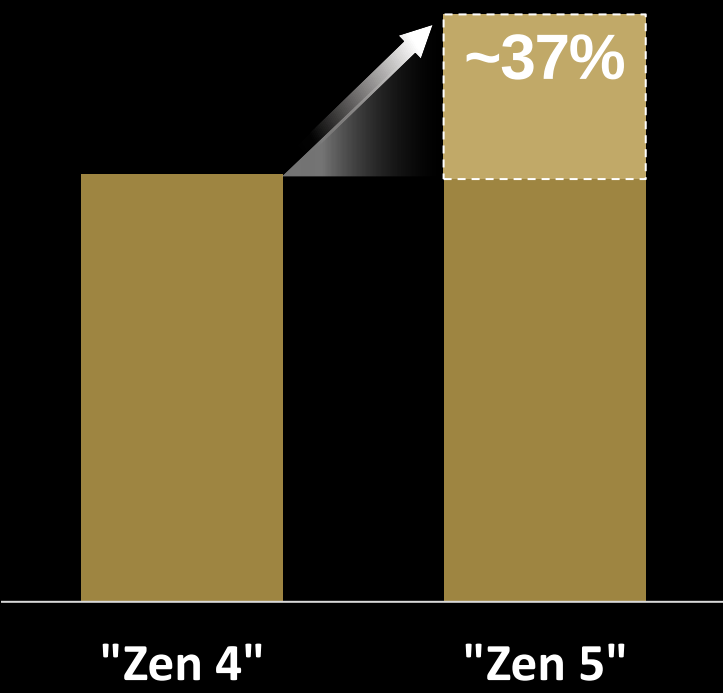
AMD EPYC™ Processors Generational Improvements

“Zen 5” Delivering Exceptional IPC Uplift for Server CPUs

Geomean of 36 Enterprise & Cloud Server Workloads
(Fixed Frequency, 12+1 CCD)

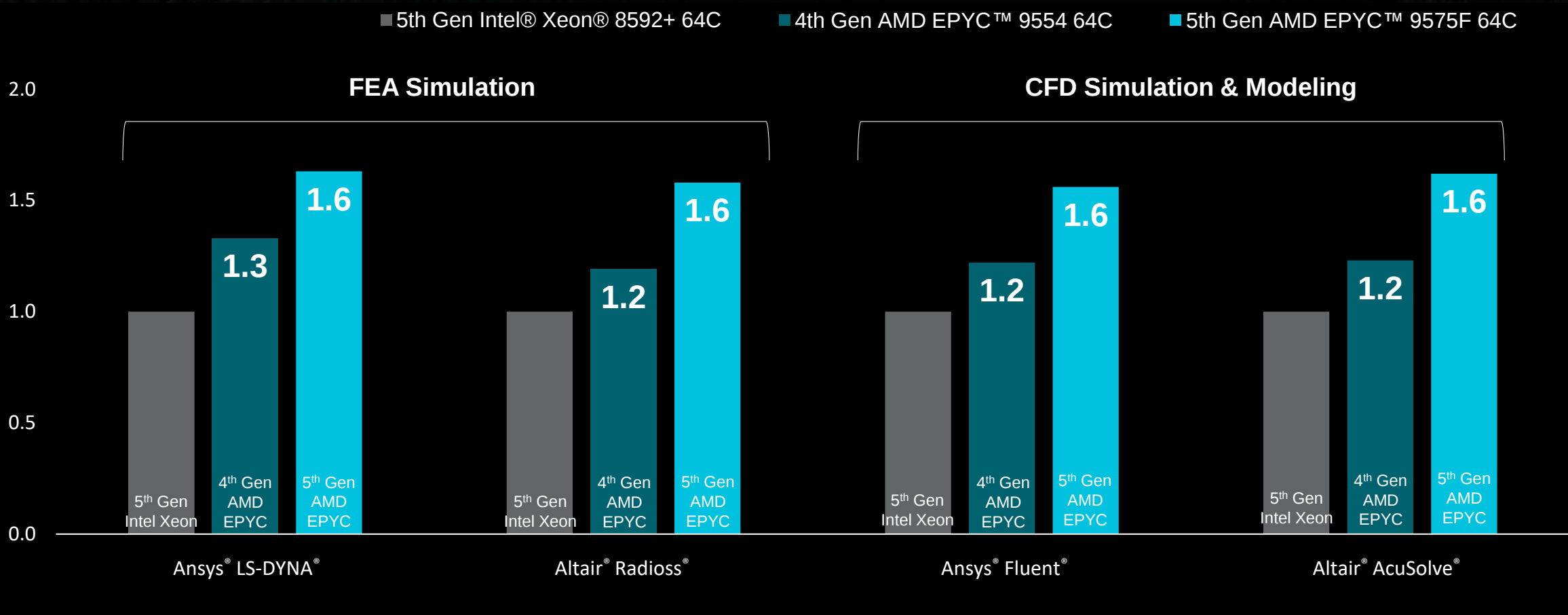


Geomean of 24 AI & HPC Server Workloads
(Fixed Frequency, 12+1 CCD)



Enterprise HPC Performance Leadership

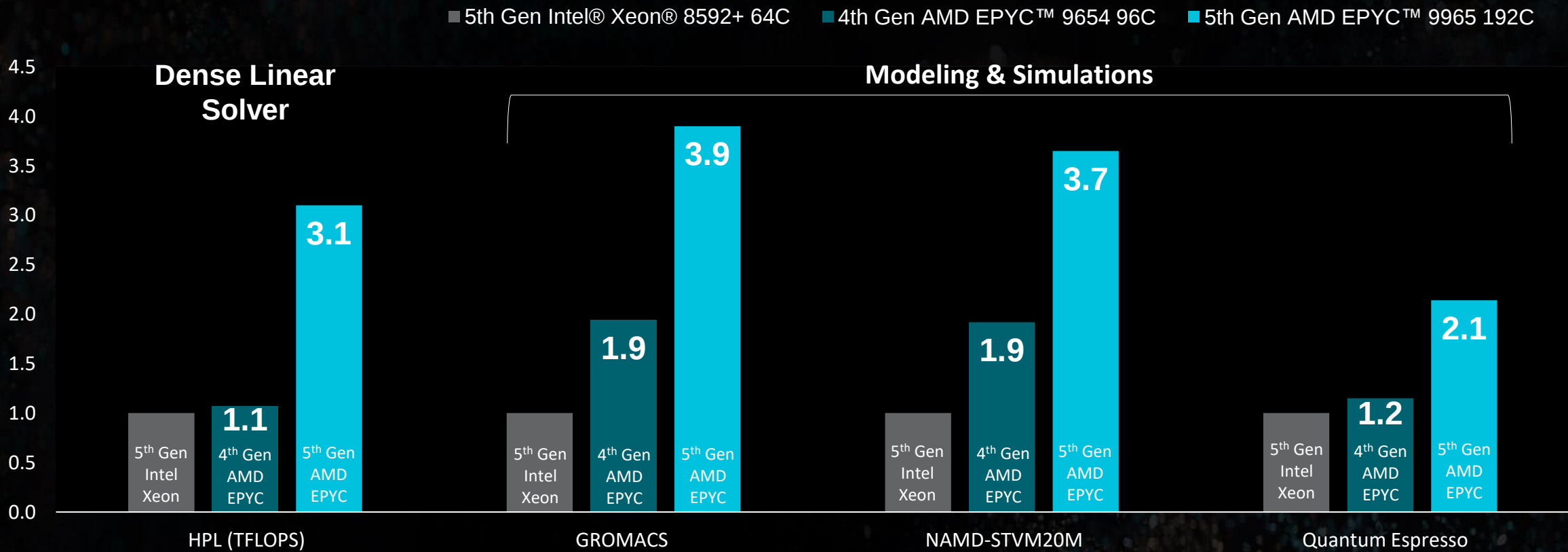
Improved Time to Insight – Up to **1.6X** at 64 cores



As of 10/10/2024. See endnotes 9xx5-028, 029, 030, 031, 032, 033, 034A, 035A

Opensource HPC Performance Leadership

Solving the Most Challenging Problems Quickly - Up to **3.9X**



As of 10/10/2024. See endnotes 9xx5-022, 023, 024, 038, 039

