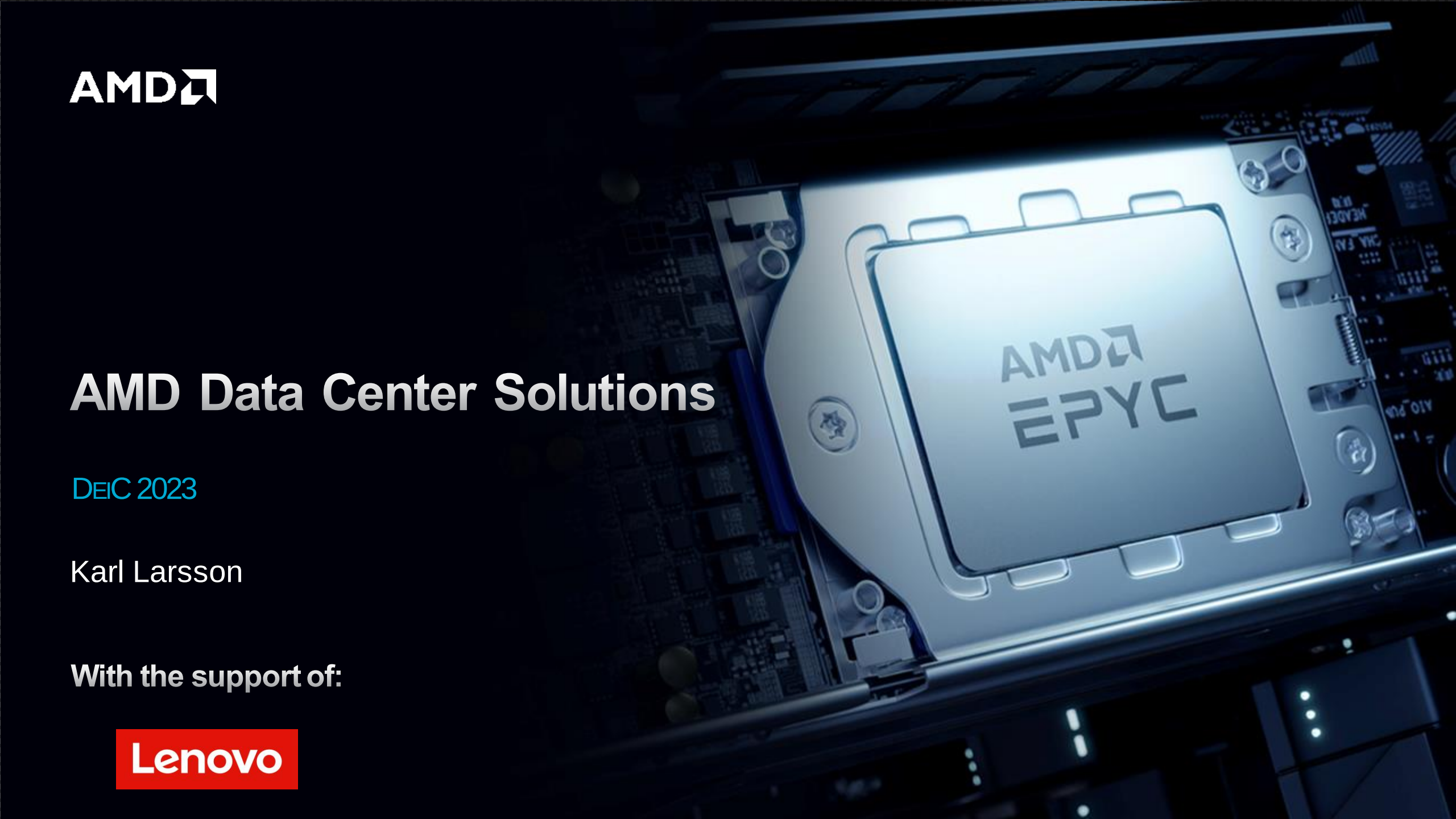



# AMD Data Center Solutions

DEIC 2023

Karl Larsson

**With the support of:**

Lenovo

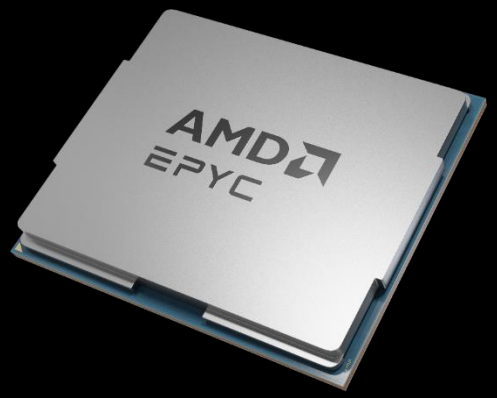
# AGENDA

- AMD Data Center Portfolio & Strategy
- EPYC CPU Portfolio
- MI300 series GPUs
- Sustainability Strategy, Goals & Results

Lenovo

AMD
together we advance_

# MODERN DATA CENTERS NEED
# WORKLOAD-OPTIMIZED ENGINES




Server CPUs

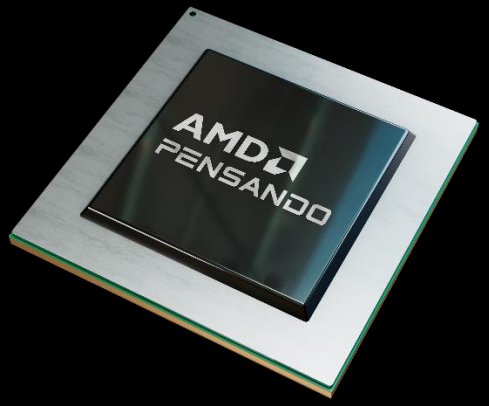
AMD EPYC


AI Accelerators

AMD INSTINCT AMD ALVEO


FPGAs and Adaptive SoCs

AMD ALVEO AMD VERSAL




SmartNICs and DPUs

AMD ALVEO AMD PENSANDO

Lenovo

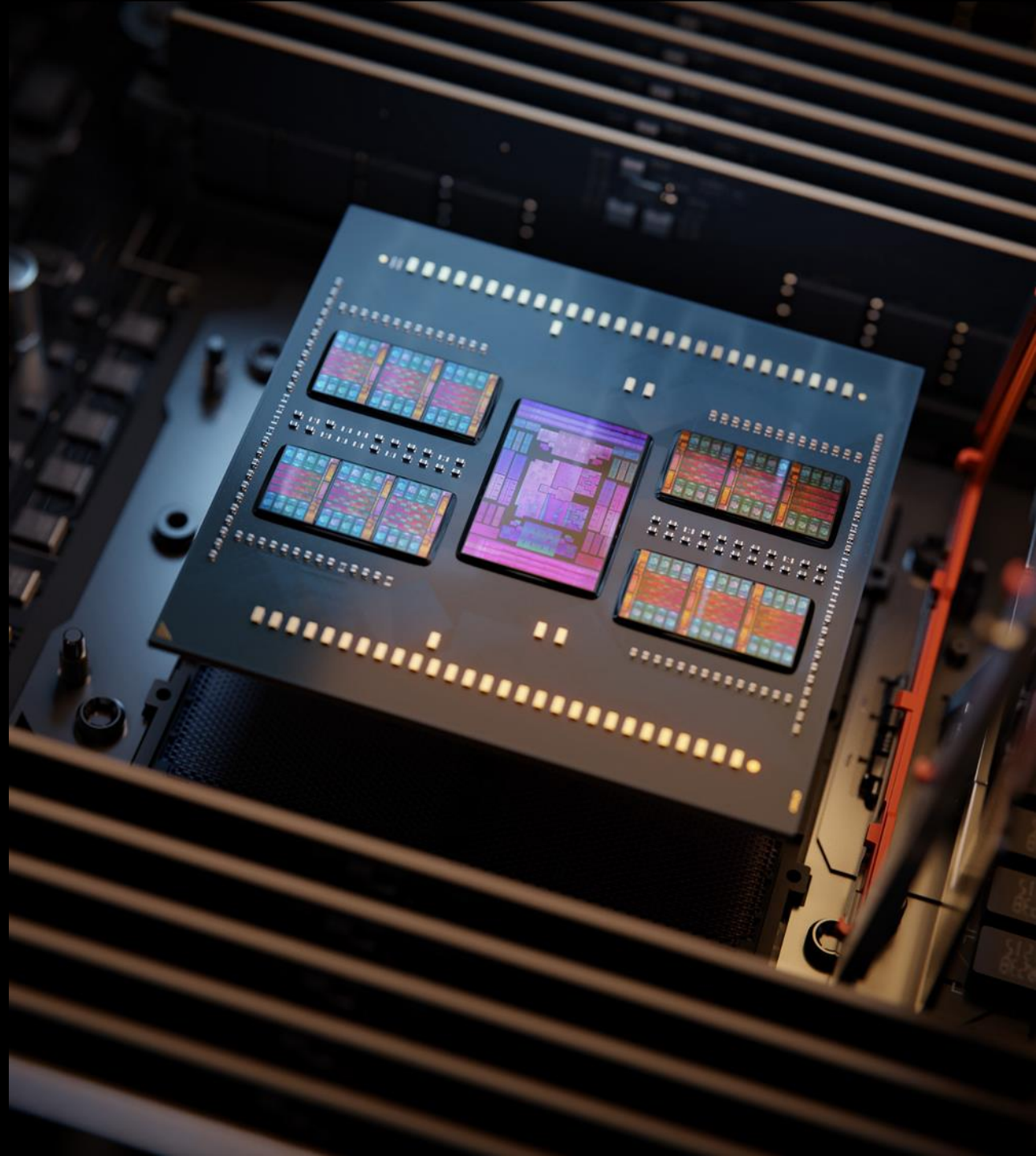
AMD
together we advance_

AMD

# CHIPLET ARCHITECTURE LEADERSHIP

## Scaling Beyond Moore's Law

- Modular, configurable design
- Leading process nodes, advanced packaging, 3D stacking
- Accelerated performance gains
- Power and cost efficiencies

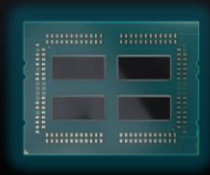


Lenovo

# AMD EPYC™ CPU Journey

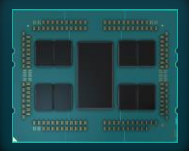
## Four Generations of On-Time Execution…Continues!

1st Gen
AMD EPYC™



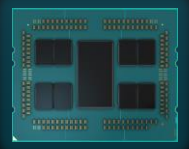
General Purpose
"Naples"

2nd Gen
AMD EPYC™

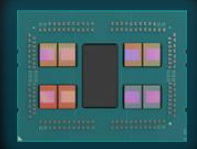


General Purpose
"Rome"

3rd Gen
AMD EPYC™

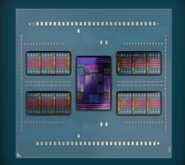


General Purpose
"Milan"

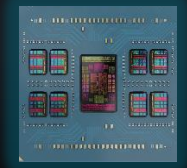


Technical
"Milan-X"

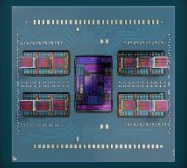
4th Gen
AMD EPYC™



General Purpose
"Genoa"



Cloud Native
"Bergamo"



Technical
"Genoa-X"

EPYC 9004
Common Socket and Platform

Intelligent Edge
"Siena"

EPYC 8004

2017 → 2023

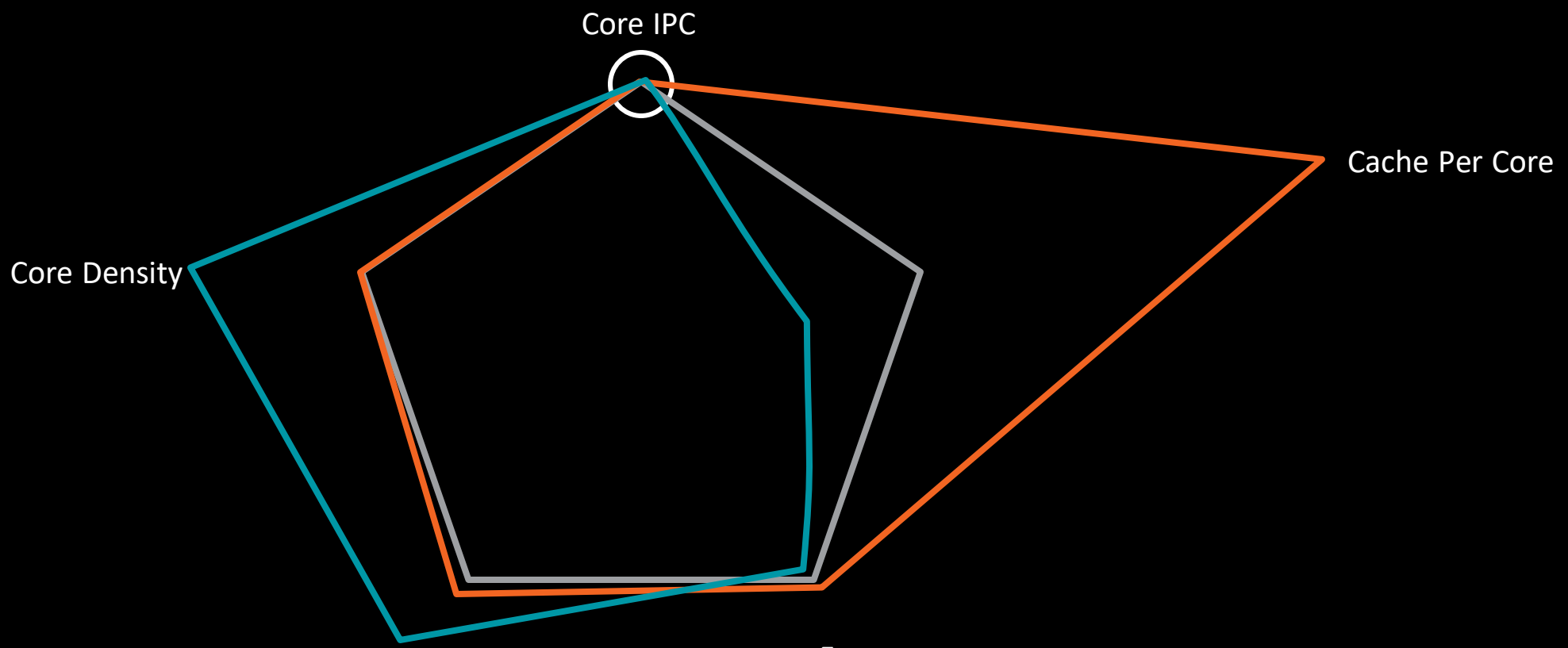
Lenovo



AMD
together we advance_

# 4th Gen AMD EPYC™ CPU

## Optimized For Workload From A Common Architecture




Lenovo

AMD
together we advance_

# AMD EPYC™ 9004 Series Processor

## All-in Feature Set support

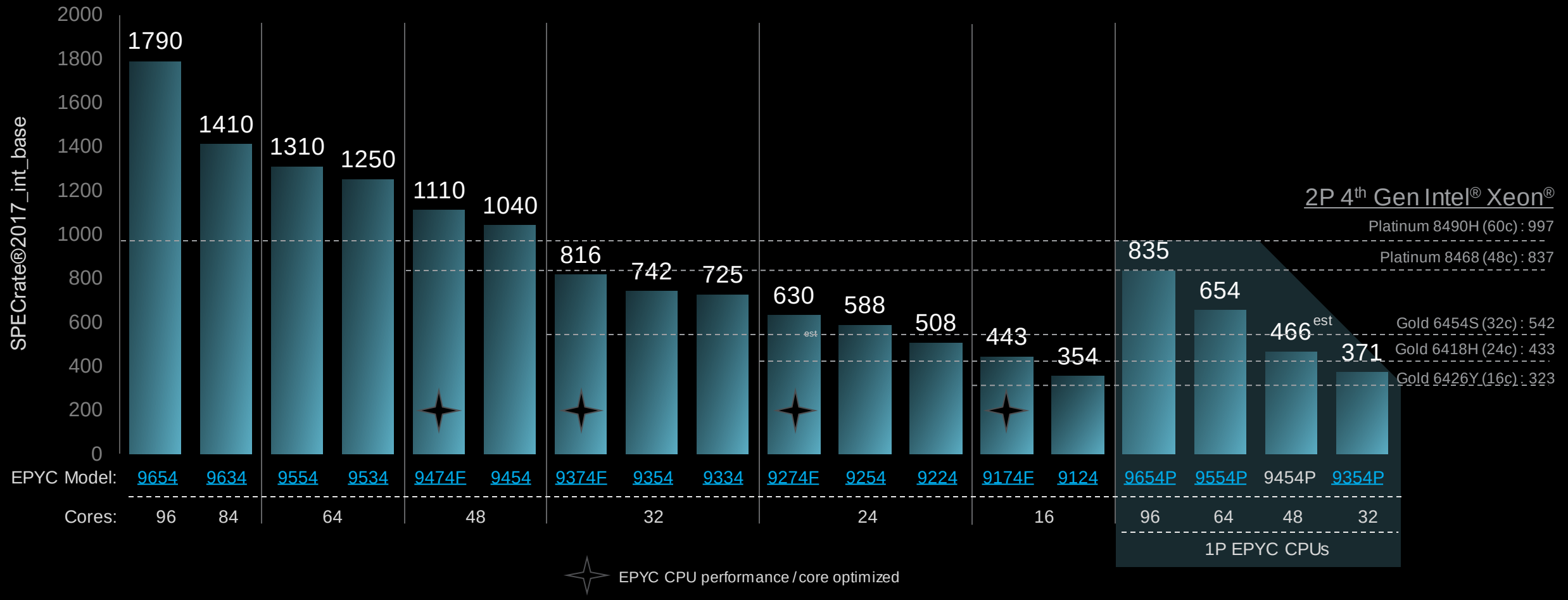
- 12 Channels of DDR5-4800
- Up to 6TB DDR5 memory capacity
- 128 lanes PCIe® 5
- 64 lanes CXL 1.1+
- AVX-512 ISA, SMT & core frequency boost
- AMD Infinity Fabric™
- AMD Infinity Guard

| Cores | AMD EPYC | Base/Boost* (up to GHz) | Default TDP (w) | cTDP (w) |
|---|---|---|---|---|
| 128 cores | 9754 | 2.25/3.10 | 360w | 320-400w |
| 112 cores | 9734 | 2.20/3.00 | 340w | 320-400w |
| 96 cores | **9684X** | 2.55/3.70 | 400w | 320-400w |
| 96 cores | 9654/P | 2.40/3.70 | 360w | 320-400w |
| 84 cores | 9634 | 2.25/3.70 | 290w | 240-300w |
| 64 cores | 9554/P | 3.10/3.75 | 360w | 320-400w |
| | 9534 | 2.45/3.70 | 280w | 240-300w |
| 48 cores | **9474F** | 3.60/4.10 | 360w | 320-400w |
| | 9454/P | 2.75/3.80 | 290w | 240-300w |
| 32 cores | **9384X** | 3.10/3.90 | 320w | 320-400w |
| | **9374F** | 3.85/4.30 | 320w | 320-400w |
| | 9354/P | 3.25/3.80 | 280w | 240-300w |
| | 9334 | 2.70/3.90 | 210w | 200-240w |
| 24 cores | **9274F** | 4.05/4.30 | 320w | 320-400w |
| | 9254 | 2.90/4.15 | 200w | 200-240w |
| | 9224 | 2.50/3.70 | 200w | 200-240w |
| 16 cores | **9184X** | 3.55/4.20 | 320w | 320-400w |
| | **9174F** | 4.10/4.40 | 320w | 320-400w |
| | 9124 | 3.00/3.70 | 200w | 200-240w |

*SEE ENDNOTES: EPYC-018, GD-183

# EPYC PERFORMANCE TO FIT YOUR NEEDS

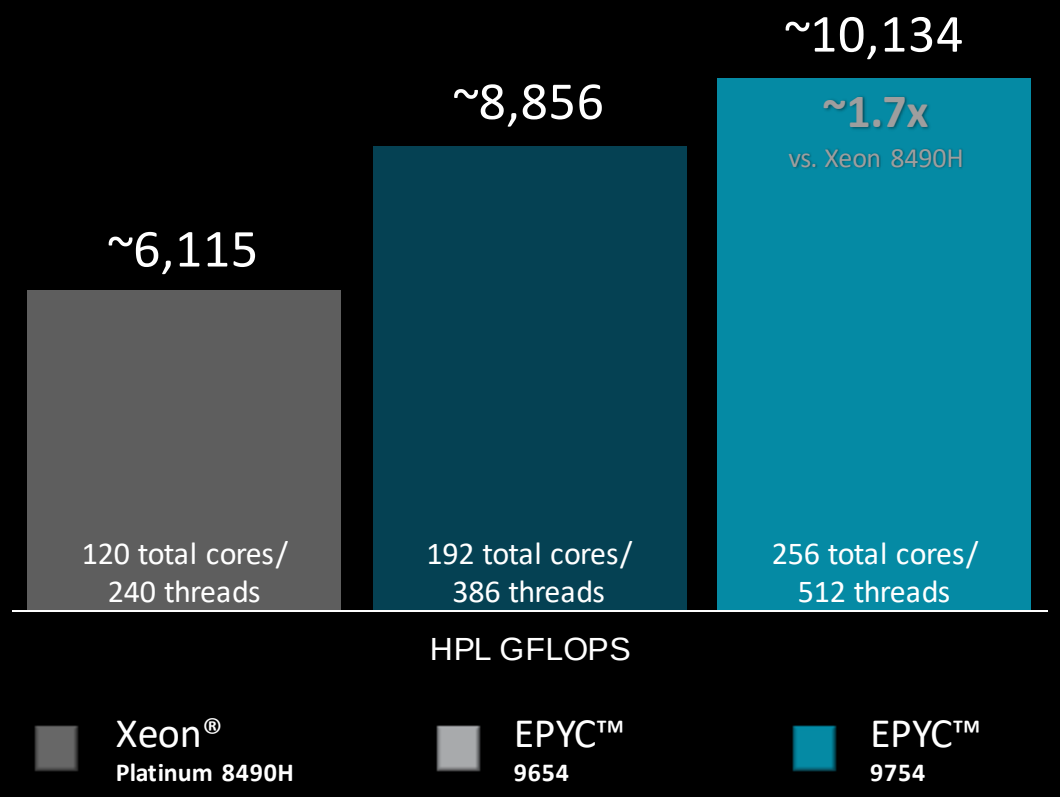
## EPYC™ 9004 Series – Performance That Scales



# Highest FLOPS to Solve the Biggest HPC Problems

## running 2P servers with 128C EPYC™ 9754 and 96C EPYC 9654 vs. 60C Xeon Platinum 8490H

Matrix Multiplication




- High Performance Linpack (HPL) is used to measure supercomputer performance – proxy for compute-bound applications like life sciences
- 128C EPYC 9754 score delivers up to
  - **~14%** more GFLOPS vs. 96C EPYC 9654
  - **~1.7x** the GFLOPS vs. 60C Xeon 8490H

Demand the best compute platform to solve the most challenging HPC problems with AMD EPYC

Lenovo

Results may vary due to factors including system configurations, software versions and BIOS settings. As of 6/13/2023, see SP5-154.

AMD
together we advance_

AMD

CDNA 3

# Next-gen AI accelerator architecture

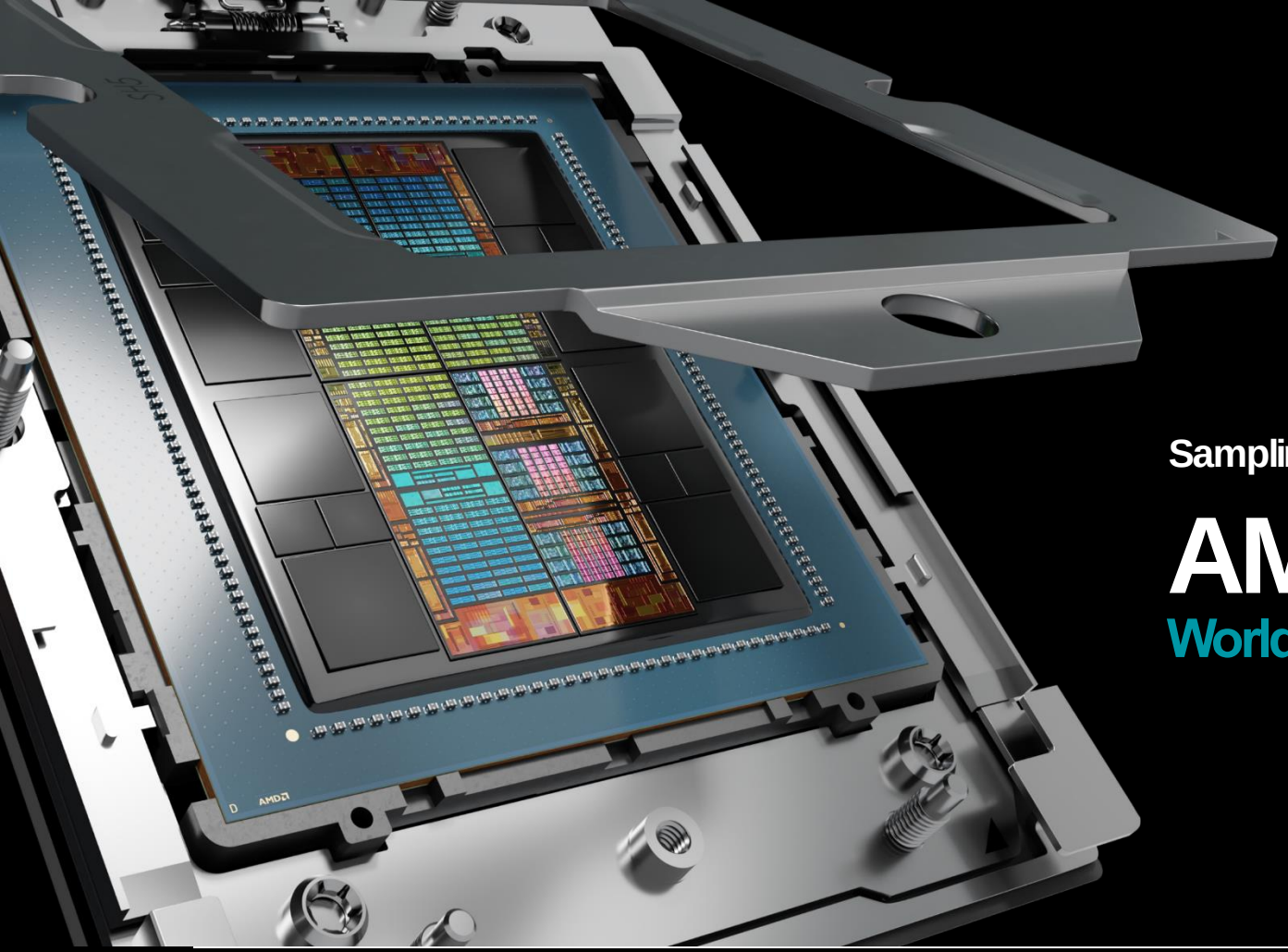
Dedicated accelerator engines for AI and HPC

3D packaging with 4th Gen AMD Infinity architecture

Optimized for performance and power efficiency

Lenovo

AMD
together we advance_



Sampling

# AMD Instinct™ MI300A

## World's first APU accelerator for AI and HPC

AMD CDNA 3 Next-Gen Accelerator Architecture

ZEN 4 24 CPU Cores

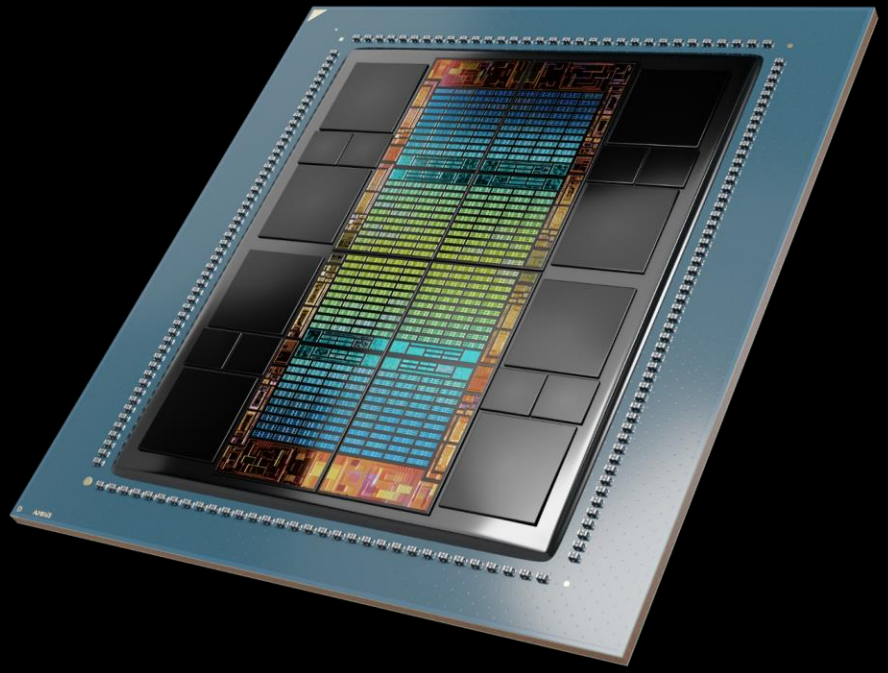
**128** GB HBM3

**5**nm and **6**nm Process Technology

**Shared Memory** CPU + GPU

Lenovo

AMD together we advance_



# AMD Instinct™ MI300X

## Leadership generative AI accelerator

**Up to 2.4x** HBM density compared to Nvidia H100

**Up to 1.6x** HBM bandwidth compared to Nvidia H100

| AMD CDNA 3 | 192 GB HBM3 | 5.2 TB/s Memory Bandwidth | 896 GB/s Infinity Fabric™ Bandwidth | 153 B Transistors |
|---|---|---|---|---|

Lenovo

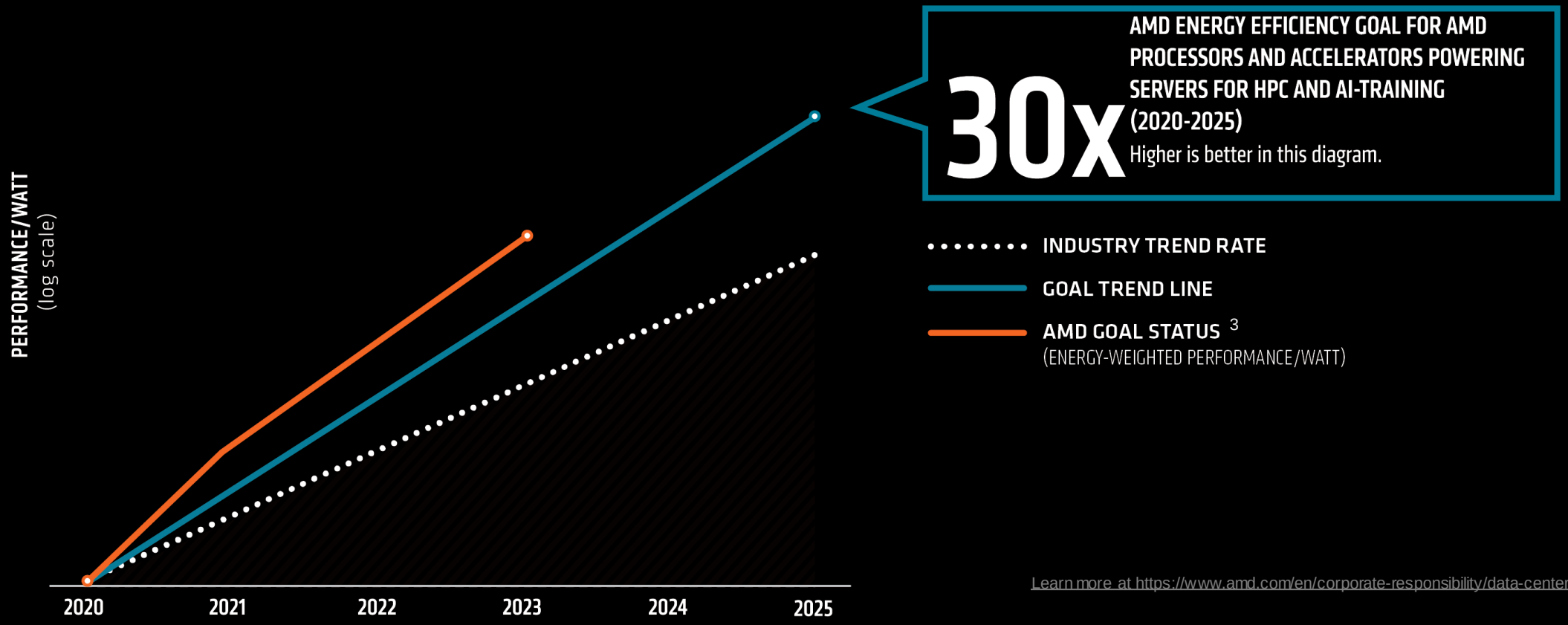
AMD together we advance_

See Endnotes: MI300-05

# Advancing data center sustainability

The AMD “30x25” goal is to deliver 30x more energy efficiency for our acerated compute nodes powering servers for AI-training and HPC (2020-2025).[2] The goal represents:

- 2.5x acceleration of the industry trends from 2015-2020 (measured by worldwide energy consumption for these computing segments)
- 97% reduction in energy use per computation from 2020-2025




Learn more at https://www.amd.com/en/corporate-responsibility/data-center-sustainability

Lenovo

AMD
together we advance_

# Advancing environmental sustainability

## Enabling innovative solutions


### Vestas

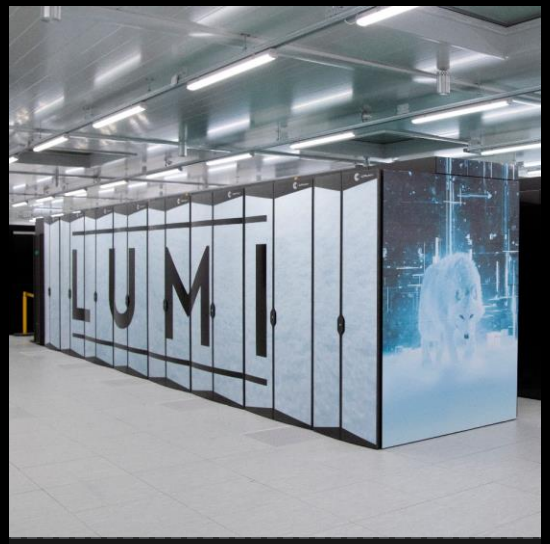
Optimizing wind turbine orchestration to reduce power lost from wake turbulence

[LINK]


### Accelerate Wind

Increasing modeling performance for faster wind turbine design development

[LINK]




### Lumi

Creating a “digital twin” of the earth to better understand and adapt to climate change

[LINK]


### KTH

Optimizing air and sea transport as well as material efficiency for solar powered systems

[LINK]

Lenovo

AMD
together we advance_

# Aarhus University reference case powered by AMD & Lenovo

60 nodes (30 trays) of SD665 V3 with 2 x 9654 CPUs

Lenovo

AMD
together we advance_

# Oceanbox reference case powered by AMD & Lenovo

Lenovo

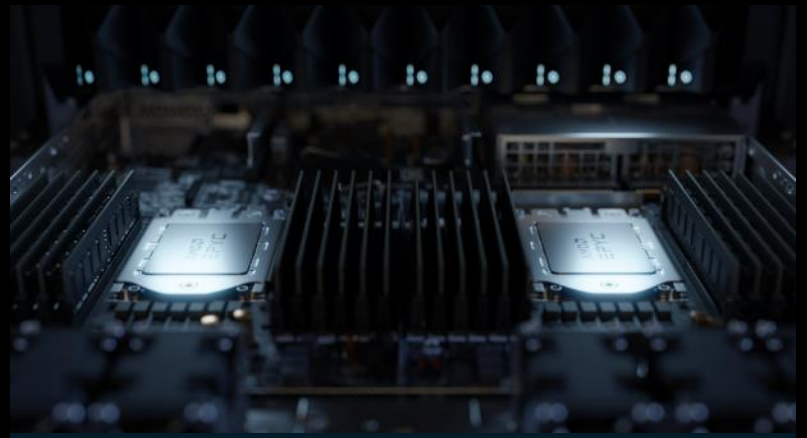
AMD
together we advance_


With the support of:

Lenovo

AMD
together we advance_

# AMD Server Strategy



Highest performing general purpose data center CPU in the world

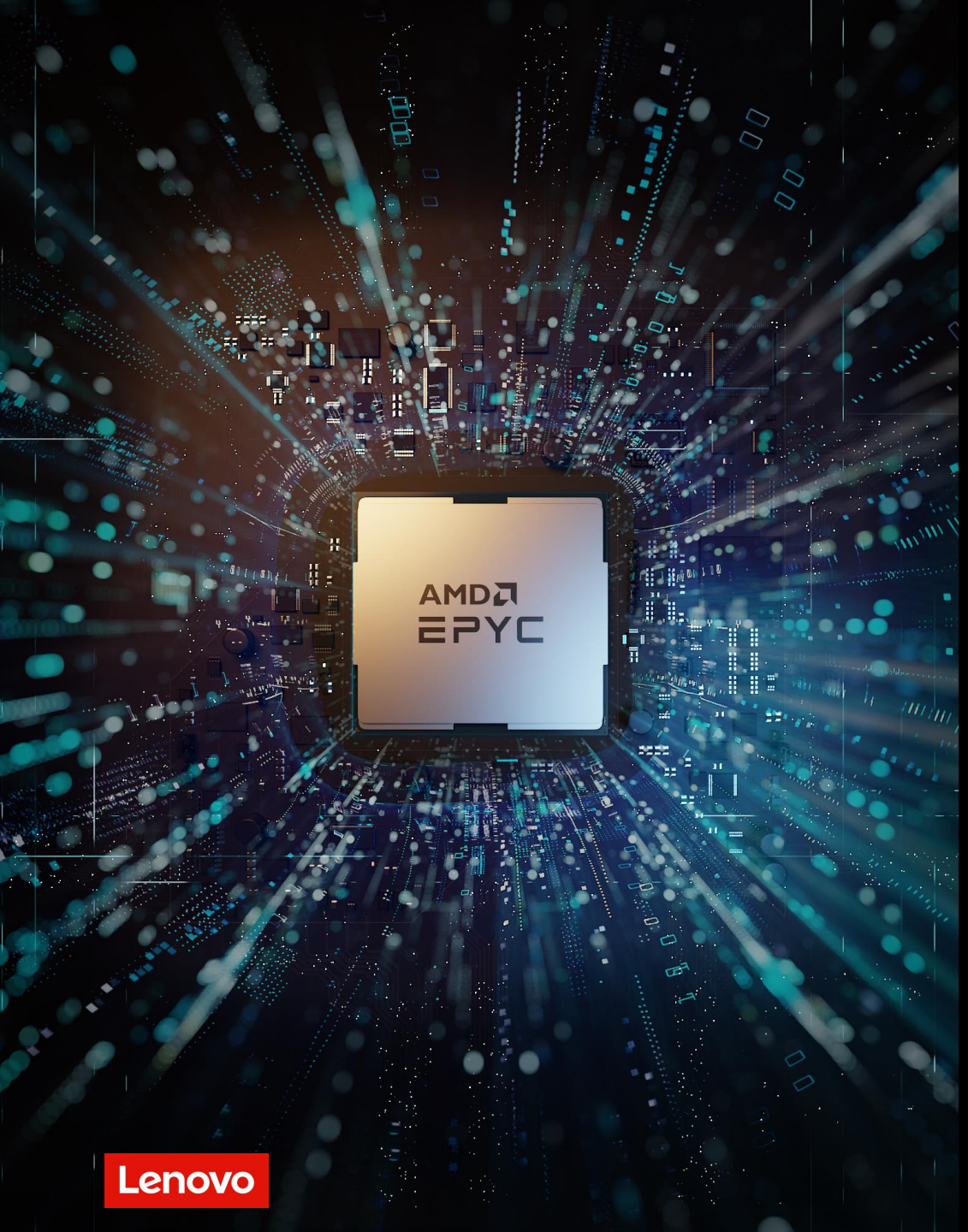

Optimized silicon for diverse workloads

Full stack solutions, ecosystem scale & partnerships to accelerate time-to-value

*ACCELERATING CUSTOMER VALUE*

Lenovo

AMD
together we advance_




# Delivering What Customers Are Asking For

World's highest performance x86 server processor

Outstanding TCO across workloads and industries

Leadership x86 energy efficiency to support sustainability goals

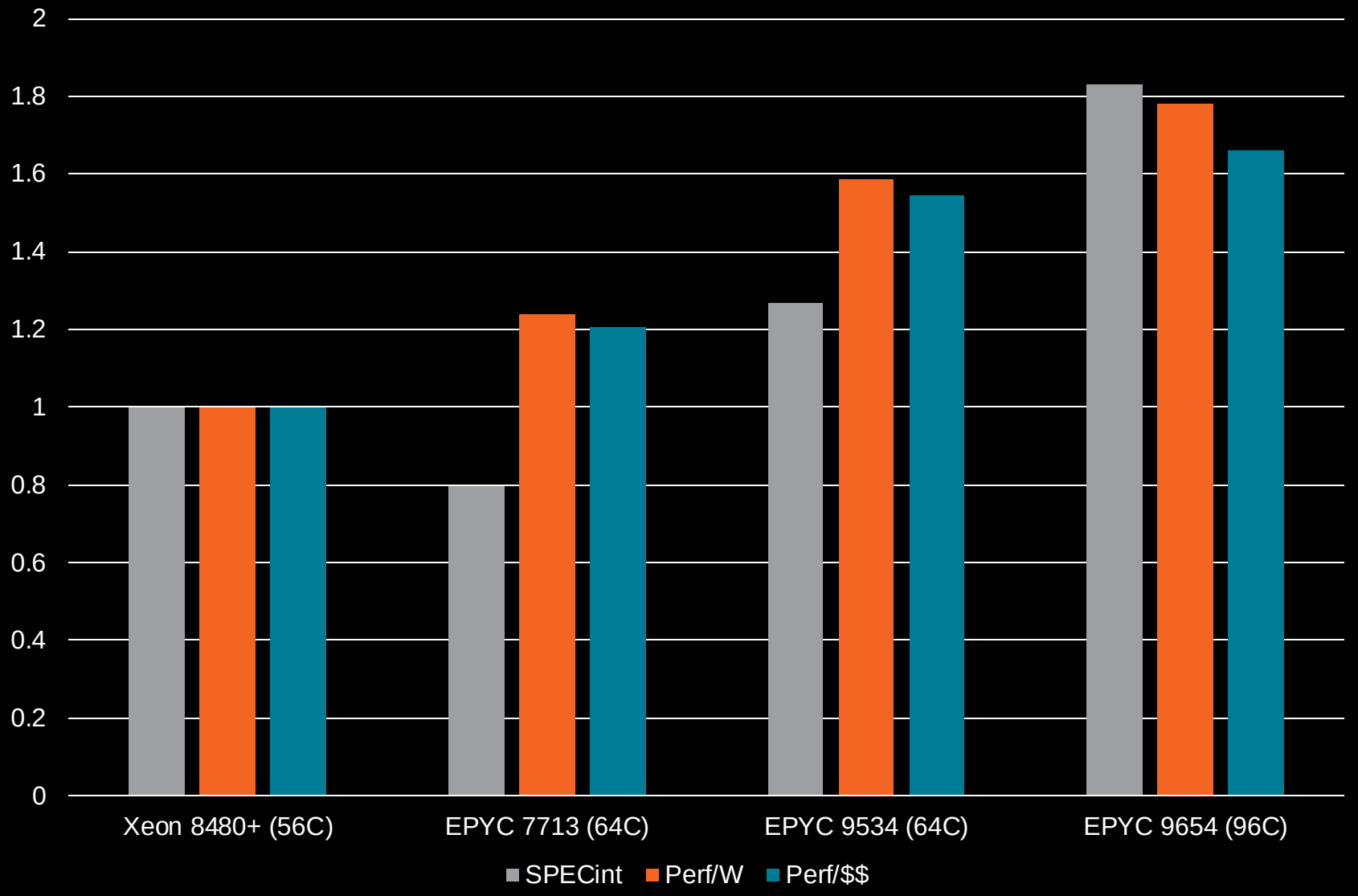
Assurance of confidential computing

Rich ecosystem of solutions

Lenovo

AMD
together we advance_

# PERF/$$ and Perf/W leadership



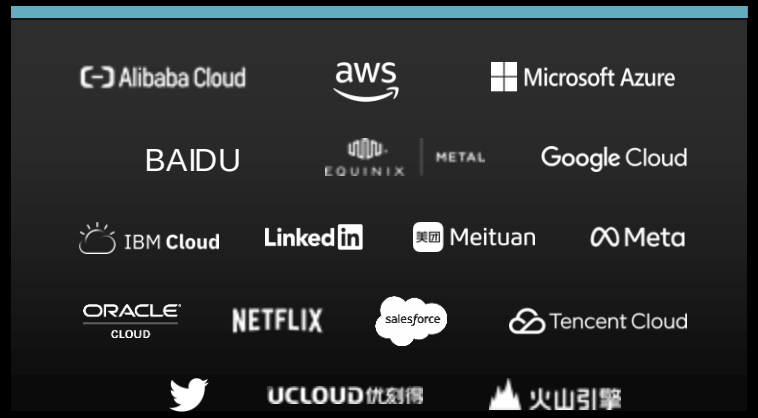

Lenovo

AMD
together we advance_

# AMD EPYC™ PROCESSOR DEPLOYMENT

## 16 STRAIGHT QUARTERS OF MARKET SHARE GROWTH

DON'T GET LEFT BEHIND



### HPC

12/20 Top Supercomputers in TOP500

7/10 Top, Most Efficient Supercomputers in Green 500

### Cloud

Powering SaaS Offerings and Internal Infrastructure of Top 10 Hyperscalers

### Enterprise

7/10 Top Financial Services Companies in North America and Europe

9/10 Top Automotive Companies

Lenovo


Top500 and Green500 lists are the latest as of June 2023 @ top500.org
Top 10 Financial and Automotive companies defined by https://www.forbes.com/lists/global2000/?sh=6d3c2e715ac0
Share gains based on Mercury Research

AMD
together we advance_

# How AMD and our partners advance environmental sustainability

Addressing environmental impacts at AMD and in our supply chain

Innovating on collaborative solutions to address environmental challenges

Advancing environmental performance for IT users

Lenovo

AMD
together we advance_

- 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-183
- AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard.

Lenovo

AMD
together we advance_

# Endnotes

- MI300-005: Calculations conducted by AMD Performance Labs as of May 17, 2023, for the AMD Instinct™ MI300X OAM accelerator 750W (192 GB HBM3) designed with AMD CDNA™ 3 5nm FinFet process technology resulted in 192 GB HBM3 memory capacity and 5.218 TFLOPS sustained peak memory bandwidth performance. MI300X memory bus interface is 8,192 and memory data rate is 5.6 Gbps for total sustained peak memory bandwidth of 5.218 TB/s (8,192 bits memory bus interface * 5.6 Gbps memory data rate/8)*0.91 delivered adjustment. The highest published results on the NVidia Hopper H100 (80GB) SXM GPU accelerator resulted in 80GB HBM3 memory capacity and 3.35 TB/s GPU memory bandwidth performance.

- MI300-08K - Measurements by internal AMD Performance Labs as of June 2, 2023 on current specifications and/or internal engineering calculations. Large Language Model (LLM) run comparisons with FP16 precision to determine the minimum number of GPUs needed to run the Falcon (40B parameters); GPT-3 (175 Billion parameters), PaLM 2 (340 Billion parameters); PaLM (540 Billion parameters) models. Calculated estimates based on GPU-only memory size versus memory required by the model at defined parameters plus 10% overhead.
  Calculations rely on published and sometimes preliminary model memory sizes. Tested result configurations: AMD Lab system consisting of 1x EPYC 9654 (96-core) CPU with 1x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator Vs. Competitve testing done on Cirrascale Cloud Services comparable instance with permission.
  Results (FP16 precision):

| Model: | Parameters | Tot Mem. Reqd | MI300X Reqd | Competition Reqd |
|---|---|---|---|---|
| Falcon-40B | 40 Billion | 88 GB | 1 Actual | 2 Actual |
| GPT-3 | 175 Billion | 385 GB | 3 Calculated | 5 Calculated |
| PaLM 2 | 340 Billion | 748 GB | 4 Calculated | 10 Calculated |
| PaLM | 540 Billion | 1188 GB | 7 Calculated | 15 Calculated |

  Calculated estimates may vary based on final model size; actual and estimates may vary due to actual overhead required and using system memory beyond that of the GPU. Server manufacturers may vary configuration offerings yielding different results.

Lenovo

AMD
together we advance_

# DISCLAIMER AND TRADEMARKS



Lenovo

AMD
together we advance_