AMD ENTERPRISE TECHNOLOGY

7 WAYS

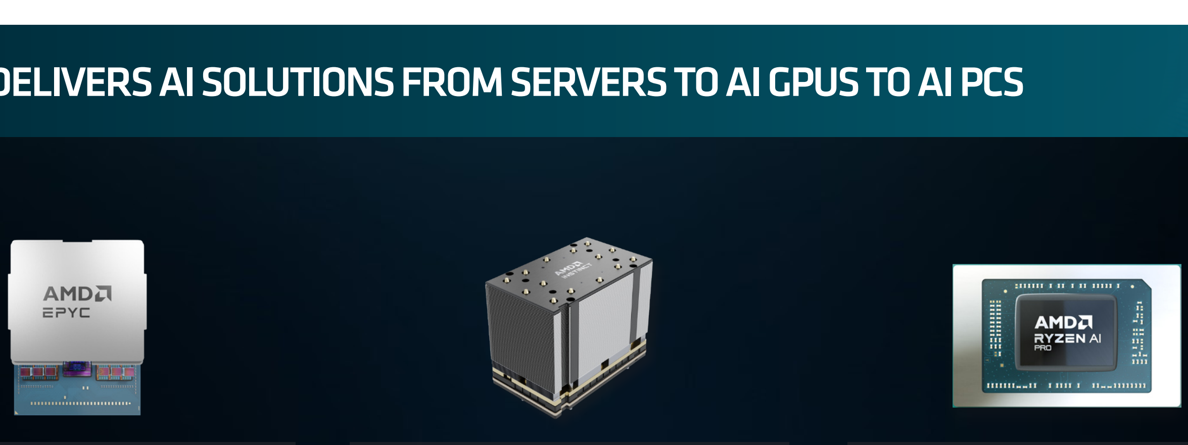
AMD ENABLES IT TRANSFORMATION IN THE ERA OF AI

1 AMD DELIVERS THE BROADEST TECHNOLOGY PORTFOLIO TO THE DATA CENTER

AMD DATA CENTER SOLUTIONS

AMD EPYC	AMD INSTINCT	AMD ALVEO AMD PENSANDO	AMD VERSAL AMD ALVEO
Server CPU Family	AI & HPC Accelerator	Networking & DPU	FPGA & Adaptive SOC

2 AMD DELIVERS AI SOLUTIONS FROM SERVERS TO AI GPUS TO AI PCS



AMD EPYC	AMD INSTINCT	AMD RYZEN PRO
General Computing	Generative AI Acceleration	Enterprise AI PCs
Leadership performance & energy efficiency provides a foundation for data center consolidation & security	Leadership generative AI performance supportingmajor AI frameworks, based on open standards	1st dedicated AI engine for Enterprise PC, with leadership performance, battery life, and security features

3 ADDRESSING THE SPECTRUM OF AI ON-PREM OR IN THE CLOUD

LEADERSHIP CPU & GPU PORTFOLIO TO SOLVE AI CHALLENGES

AMD EPYC™ CPUs & AMD Instinct™ accelerators excel in AI workloads with different deployment scenarios

AMD EPYC

- AI Training
- Dedicated AI Deployments
- Medium to Large Models
- Large-Scale Inference

AMD EPYC™ consolidation advantages get data centers ready for AI, and can be used for smaller scale AI inference deployments

AMD INSTINCT

- Mixed workload Inference
- Small to Medium Models
- Batch/Small Scale Inference

AMD Instinct™ MI300 offers leadership generative AI performance supporting larger AI models than the competition

4 DATA CENTER CONSOLIDATION POWERED BY AMD EPYC™ PROCESSORS

AMD EPYC

New consolidation approaches have transformed what is achievable in the data center. These approaches can offer greater performance and high efficiency, all within the same or smaller footprint.

AMD EPYC™ consolidation advantages get data centers ready for AI, and can be used for small scale AI inference deployments.

Estimated comparisons. See endnotes SP5TCO-055, SP5TCO-056

73%	70%	65%
Fewer Servers	Fewer Racks	Less Power

Sky Lake Intel® Xeon® 6143
vs 4th Gen AMD EPYC™ 9334

68%	65%	56%
Fewer Servers	Fewer Racks	Less Power

Cascade Lake Intel® Xeon® 6242
vs 4th Gen AMD EPYC™ 9334

Target: 80,000 Integer Performance

5 AMD RYZEN™ PRO CPUS POWER THE BEST BUSINESS PCS IN THE WORLD

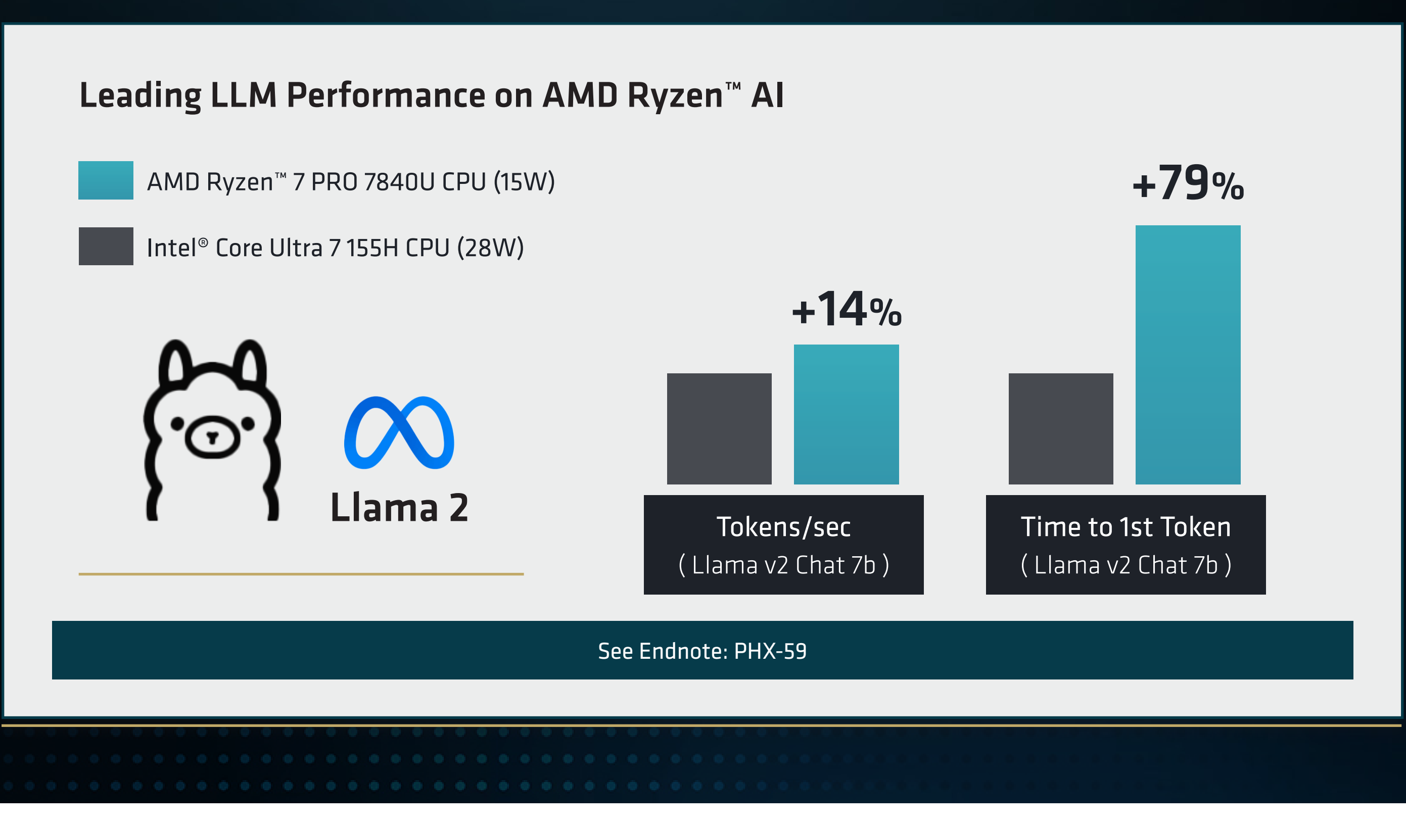
PERFORMANCE UPLIFT AND ENERGY SAVINGS ACROSS THE PC FLEET

AMD Ryzen™ AI is the leading CPU for AI PCs, and enables the workforce for AI with a wide variety of laptop, desktop, and workstation solutions from OEMs

Microsoft Teams Call (up to)	**with AMD Ryzen™ AI** (up to)
+ Procyon Office Productivity Word Benchmark	+ Procyon Office Productivity Excel Benchmark
69% Faster Performance	72% Faster Performance
+ Procyon Office Productivity Benchmark	+ Procyon Office Productivity Benchmark
55% Less Power	84% Less Power

AMD Ryzen™ PRO 8840U vs Intel® Core Ultra 7 165U on Production Platforms

See Endnotes: HWKP-26, HWKP-27, HWKP-28, HWKP-29



Leading LLM Performance on AMD Ryzen™ AI

See Endnote: PHX-59

6 A COMPLETE SOFTWARE SOLUTIONS ECOSYSTEM

AMD EPYC™ has a complete software ecosystem of solutions comparable to Intel's that is fully compatible with existing software in enterprise data centers.

HPC & AI	ALTAIR, Ansys, DASSAULT SYSTEMES, deci, NEURAL MAGIC, ONNX, PyTorch, SIEMENS, TensorFlow
Database Analytics	CLOUDERA, Couchbase, databricks, DATASTAX, elastic, MarkLogic, mongoDB, splunk>, SingleStore, VERTICA
Database	Exasol, influxdata, Microsoft SQL Server, ORACLE, PostgreSQL, redislabs, TigerGraph, SAP, cassandra, MySQL, MariaDB
OS	CANONICAL, citrix, FreeBSD, Microsoft, NUTANIX, ORACLE, Red Hat, SUSE, vmware
HCI / Orchestration	docker, kubernetes, Microsoft, NUTANIX, Red Hat, simplivity, vmware
SDS	CLOUDIAN, Excelero, Pivot3, Quobyte, ceph, StorMagic, WEKA

Fully Compatible with Existing Software: Implements x86-64 Instruction Set Developed by AMD, Adopted by Intel

ENTERPRISE-READY OPEN AI SOFTWARE

Deep engagement with AI leaders like Hugging Face, PyTorch and a broad set of AI software ecosystem enables a rich ecosystem around AMD ROCm™ and AMD Instinct™ bringing an open, performant and proven GPU solution to the market.

Hugging Face	PyTorch	ONNX RUNTIME, OpenXLA, DeepSpeed
62,000+ Models Running Nightly Fully Integrated Optimum library	Extending From "Day 0" to "Bleeding-Edge" Support	Expanding Open Source Contributions & Footprint

7 AMD INSTINCT™ ACCELERATORS SUPERCHARGE AI & HPC

Powering the most demanding AI and HPC workloads, offering exceptional compute performance, large memory density, high bandwidth memory, and support for specialized data formats.

1 Nvidia H100 HGX 640 GB HBM3 | 26.4 TB/s	1 AMD Instinct™ MI300X Platform 1.5 TB HBM3 | 42.4 TB/s		
Training & Inference	Training	Inference	
1x	1x MPT-30B	1.6x Bloom 176B	Perf/ System
1x	2x ~30B	2x ~70B	Models/ System
1x	2x ~70B vs ~30B	2x ~680B vs 290B	Max LLM Size/System

Results may vary. See endnotes: MI300-34, MI300-39, MI300-40, MI300-42

Incredible performance upgrade for GenAI training and inference

AMD Instinct™ MI300X vs Nvidia H100 vLLM (FP16) comparison	2.1x	Llama 2 | 70B Latency Improvement
AMD Instinct™ MI300X vs AMD Instinct™ MI250*	8x	Llama 2 | 70B Latency Improvement

Available from leading OEMs & CSPs

Cirrascale, DELL Technologies, Hewlett Packard Enterprise, Lenovo, Microsoft, ORACLE, SUPERMICRO

* AMD Instinct™ MI300X using AMD ROCm™ 6 & AMD Instinct™ MI250 using AMD ROCm™ 5

Results may vary. See endnotes: MI300-33, MI300-38A

THE LARGEST AND MOST DISCERNING HYPERSCALE DATA CENTER CUSTOMERS CHOOSE AMD EPYC™ CPUS TO POWER INTERNAL WORKLOADS SERVING BILLIONS OF USERS

HWKP-26. Testing as of 3/26/24 by AMD Performance Labs on a HP EliteBook 845 G11 with an AMD Ryzen™ 7 PRO 8840U processor @15W, integrated Radeon™ 780M graphics, 32GB RAM (2X16GB) 2800MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, 16GB RAM (2X8GB) 29866.7 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Arc Graphics, 16GB RAM (2X8GB) 33600.0 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional. The following applications were tested in Balanced Mode: Teams + Procyon Office Productivity, Teams + Procyon Office Productivity Excel, Teams + Procyon Office Productivity Outlook, Teams + Procyon Office Productivity Power Point, Teams + Procyon Office Productivity Word, Composite Geomean Score. Each Microsoft Teams call consists of 9 participants (3X3). Laptop manufactures may vary configurations yielding different results.

HWKP-27. Testing as of 3/26/24 by AMD Performance Labs on a HP EliteBook 845 G11 with an AMD Ryzen™ 7 PRO 8840U processor @15W, integrated Radeon™ 780M graphics, 32GB RAM (2X16GB) 2800MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, 16GB RAM (2X8GB) 29866.7 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Arc Graphics, 16GB RAM (2X8GB) 33600.0 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs a Dell XPS 9440 with Intel Core Ultra 7 155H processor @28W, Intel Arc Graphics, 16GB RAM (2X8GB) 6400MHz, 1TB NVMe SSD, 69Wh battery, Microsoft Windows 11 Professional. All systems run with the camera and background blur ON, in Best Power Efficiency mode using the following applications: Microsoft Teams + Procyon Office Productivity Overall benchmark measuring Wall power consumed (watts). Each Microsoft Teams call consists of 9 participants (3X3). Laptop manufacturers may vary configurations yielding different results

HWKP-28. Testing as of 3/26/24 by AMD Performance Labs on a HP EliteBook 845 G11 with an AMD Ryzen™ 7 PRO 8840U processor @15W, integrated Radeon™ 780M graphics, 32GB RAM (2X16GB) 2800MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, 16GB RAM (2X8GB) 29866.7 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Arc Graphics, 16GB RAM (2X8GB) 33600.0 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional. The following applications were tested in Balanced Mode: Microsoft Teams with AI enabled (Windows Studio Effects) + Procyon Office Productivity, Procyon Office Productivity Excel, Procyon Office Productivity Outlook, Procyon Office Productivity Power Point, Procyon Office Productivity Word, Composite Geomean Score. Each Microsoft Teams call consists of 9 participants (3X3). Laptop manufactures may vary configurations yielding different results

HWKP-29. Testing as of 3/26/24 by AMD Performance Labs on a HP EliteBook 845 G11 with an AMD Ryzen™ 7 PRO 8840U processor @15W, integrated Radeon™ 780M graphics, 32GB RAM (2X16GB) 2800MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165U processor @15W (vPro enabled), Intel Iris Xe Graphics, 16GB RAM (2X8GB) 29866.7 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs. a Dell Latitude 7450 with Intel Core Ultra 7 165H processor @28W (vPro enabled), Intel Arc Graphics, 16GB RAM (2X8GB) 33600.0 MHz, 512GB NVMe SSD, Microsoft Windows 11 Professional vs a Dell XPS 9440 with Intel Core Ultra 7 155H processor @28W, Intel Arc Graphics, 16GB RAM (2X8GB) 6400MHz, 1TB NVMe SSD, 69Wh battery, Microsoft Windows 11 Professional. All systems run in Best Power Efficiency mode using the following applications: Microsoft Teams with AI enabled (All Windows Studio Effects enabled) + Procyon Office Productivity Overall benchmark measuring Wall power consumed (watts). Each Microsoft Teams call consists of 9 participants (3X3). Laptop manufacturers may vary configurations yielding different results

SP5TCO-055: This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The Bare Metal Server Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool – v9.37 Pro Refresh, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 80,000 units of integer performance based on the published scores for these specific Intel Xeon and AMD EPYC CPU based servers as of June 1, 2023. This estimation reflects a 3-year time frame with a PUE of 1.7 and a power US power cost of $0.128 / kWh. This analysis compares a 2P AMD 32 core EPYC 9334 CPU powered server with a SPECrate®2017_int_base a score of 725, https://spec.org/cpu2017/results/res2023q1/cpu2017-20230102-33282.pdf; to a 2P Intel Xeon 16 core Gold_6143 based server with a SPECrate®2017_int_base score of 197, https://spec.org/cpu2017/results/res2017q4/cpu2017-20171114-00863.pdf. Due to the wide variation of costs for real estate or admins, this TCO does not include their costs in this analysis. New AMD powered server OpEx consists of power only. The OpEx for the legacy install base of servers with Intel CPUs consists of power plus the extended warranty costs. Cost to extend the server warranty support is calculated to be 20% annually of the initial purchase price which is calculated using 2023 costs. Using this and the power costs mean that the AMD solution for a 3yr TCO is more than $2.5 million less (62% less) and has a $1.2 million or 93% lower annual OpEx.Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. For more detail see https://www.amd.co/en/claims/epyc4

SP5TCO-056: This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. The Bare Metal Server Greenhouse Gas Emissions TCO (total cost of ownership) Estimator Tool – v9.37 Pro Refresh, compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 80,000 units of integer performance based on the published scores for these specific Intel Xeon and AMD EPYC CPU based servers as of June 1, 2023. This estimation reflects a 3-year time frame with a PUE of 1.7 and a power US power cost of $0.128 / kWh. This analysis compares a 2P AMD 32 core EPYC 9334 CPU powered server with a SPECrate®2017_int_base a score of 725, https://spec.org/cpu2017/results/res2023q1/cpu2017-20230102-33282.pdf; to a 2P Intel Xeon 16 core Gold_6242 based server with a SPECrate®2017_int_base score of 231, https://spec.org/cpu2017/results/res2021q3/cpu2017-20210811-28673.pdf. Due to the wide variation of costs for real estate or admins, this TCO does not include their costs in this analysis. New AMD powered server OpEx consists of power only. The OpEx for the legacy install base of servers with Intel CPUs consists of power plus the extended warranty costs. Cost to extend the server warranty support is calculated to be 20% annually of the initial purchase price which is calculated using 2023 costs. Using this and the power costs mean that the AMD solution for a 3yr TCO is more than $1 million less (41% less) and has a $777,434 or 89% lower annual OpEx. Environmental impact estimates made leveraging this data, using the Country / Region specific electricity factors from the '2020 Grid Electricity Emissions Factors v1.4 – September 2020', and the United States Environmental Protection Agency 'Greenhouse Gas Equivalencies Calculator'. https://www.amd.co/en/claims/epyc4

MI300-33: Text generated with Llama2-70b chat using input sequence length of 4096 and 32 output token comparison using custom docker container for each system based on AMD internal testing as of 11/17/2023.Configurations: 2P Intel Xeon Platinum CPU server using 4x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.Vs.2P AMD EPYC 7763 CPU server using 4x AMD Instinct™ MI250 (128 GB HBM2e, 560W) GPUs, ROCm® 5.4.3, PyTorch 2.0.0., HuggingFace Transformers 4.35.0, Ubuntu 22.04.6.4 GPUs on each system was used in this test.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI300-34: Token generation throughput using DeepSpeed Inference with the Bloom-176b model with an input sequence length of 1948 tokens, and output sequence length of 100 tokens, and a batch size tuned to yield the highest throughput on each system comparison based on AMD internal testing using custom docker container for each system as of 11/17/2023. Configurations: 2P Intel Xeon Platinum 8480C CPU powered server with 8x AMD Instinct™ MI300X 192GB 750W GPUs, pre-release build of ROCm™ 6.0, Ubuntu 22.04.2. Vs. An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 80GB 700W GPUs, CUDA 12.0, Ubuntu 22.04.3. 8 GPUs on each system were used in this test. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-38: Overall latency for text generation using the Llama2-70b chat model with vLLM comparison using custom docker container for each system based on AMD internal testing as of 11/23/2023. Sequence length of 2048 input tokens and 128 outputtokens.Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu® 22.04.2.Vs.An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1., PyTorch 2.1.0., vLLM v.02.2.2 (most recent), Ubuntu 22.04.3.Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations

MI300-39: Number of simultaneous text generating copies of the Llama2-70b chat model, using vLLM, comparison using custom docker container for each system based in AMD internal testing as of 11/26/2023. Configurations: 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.2.0, vLLM for ROCm, Ubuntu 22.04.2. Vs. An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 12.1., PyTorch 2.1.0. vLLM v.02.2.2 (most recent), Ubuntu 22.04.3. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-40: Testing completed 11/28/2023 by AMD performance lab using MosaicML vllm-foundry to fine tune the MPT-30b model for 2 epochs using the MosaicML instruct-v3 dataset and a max sequence length of 8192 tokens using custom docker container for each system. Configurations 2P Intel Xeon Platinum 8480C CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, ROCm® 6.0 pre-release, PyTorch 2.0.1, MosaicML llm-foundry pre-release, Ubuntu 22.04.2. Vs. An Nvidia DGX H100 with 2x Intel Xeon Platinum 8480CL Processors, 8x Nvidia H100 (80GB, 700W) GPUs, CUDA 11.8, PyTorch 2.0.1., MosaicML llm-foundry, Ubuntu 22.04.3. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations.

MI300-42: Measurements by internal AMD Performance Labs as of December 1, 2023 on current specifications and/or internal engineering calculations. Inference and training Large Language Model (LLM) run comparisons with FP16 precision to determine the largest Large Language model size that is expected to run on the 8x AMD Instinct™ MI300X (192GB) accelerator platform and on the Nvidia 8x H100 (80GB) GPUs DGX platform. 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. Multiple LLMs and parameter sizes were analyzed. Max size determined by memory capacity of 8x platform. Configurations: 8x AMD Instinct™ MI300X (192GB HBM3, OAM Module) 750W accelerator at 2,100 MHz peak boost engine clock designed with 3rd Gen AMD CDNA™ 3 5nm FinFET process technology. Vs. 8x Nvidia HGX H100 (80GB HBM3, SXM5) platform Nvidia memory specification at https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet.

PHX-59. Testing as of Feb 2023 by AMD. Sustained performance average of multiple runs with specimen prompt "Write me a story about an orange cat called mr whiskers". All tests conducted on LM Studio 0.2.16. Performance may vary. Market price retrieved on 3/4/2023 (Amazon, US). Phoenix: HP Pavilion Plus Laptop 14-ey0xxx, Ryzen 7 7840U 15W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Adrenalin Driver 24.2.1. Meteor Lake: Acer Swift SFG14-72T, Intel Core Ultra 7 155H 28W TDP, 16GB LPDDR5 6400, Windows 23H2 22631.3155, Driver 31.0.101.5333.

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: (a) AMD Ryzen 7040 and 8040 Series processors except Ryzen 5 7540U, Ryzen 5 8540U, Ryzen 3 7440U, and Ryzen 3 8440U processors; and (b) all AMD Ryzen 8000G Series desktop processors except the Ryzen 5 8500G/GE and Ryzen 3 8300G/GE. Please check with your system manufacturer for feature availability prior to purchase. GD-220b.