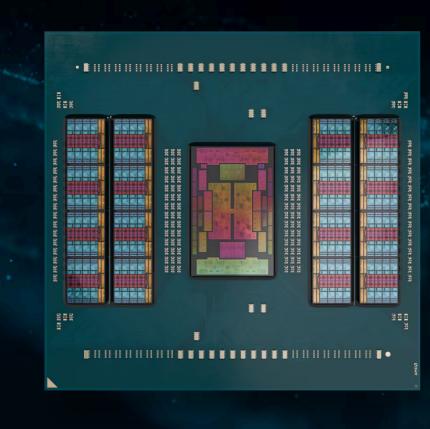


KI kann medizinische Entdeckungen beschleunigen, die Forschung revolutionieren, effizientere Städte schaffen und die Produktivität in fast jeder Branche steigern.

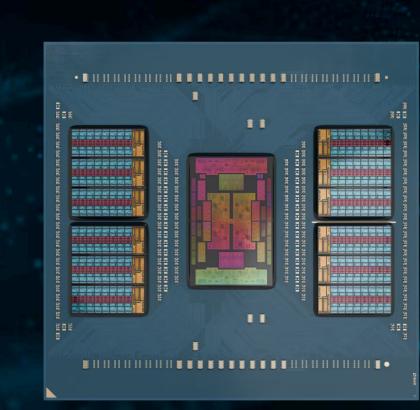
Der Erfolg mit KI hängt von der Wahl der richtigen Tools für die Aufgaben ab. Vom Chip bis zur Software. Von Rechenzentren bis zu KI-PCs. AMD entwickelt End-to-End-KI-Lösungen und treibt so Innovation im großen Maßstab voran, damit Unternehmen die Wirkung der KI-Initiativen maximieren können.

DIE BESTE CPU FÜR KI AMD EPYCTM 9005 PROZESSOREN



Bis zu 1,7-facher KI-Durchsatz

Server, die mit zwei AMD EPYC™ 9965 CPUs der 5. Generation ausgestattet sind, bieten einen bis zu 1,7-fachen KI-Durchsatz im Vergleich zur vorherigen Generation.²



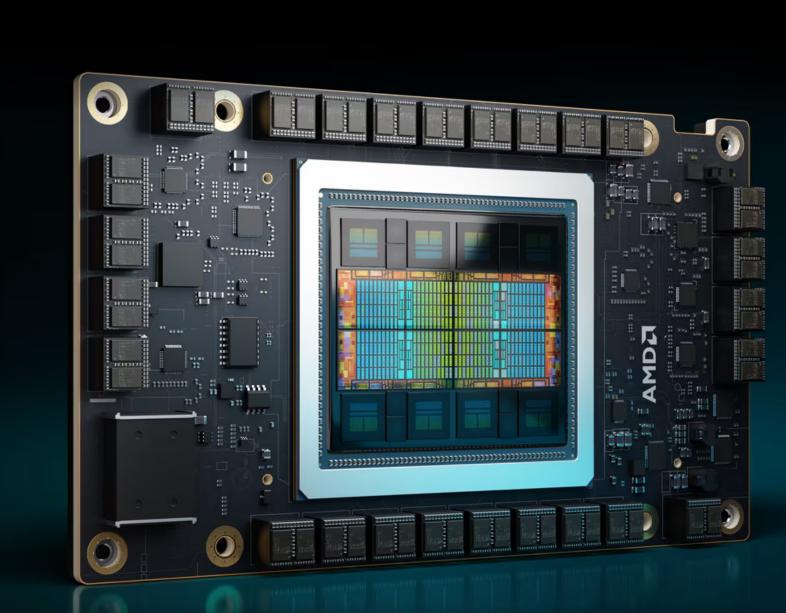
Bis zu 86 % weniger Racks

AMD EPYC™ 9005 Prozessoren können die Integer Performance alter Hardware der Konkurrenz mit bis zu 86 % weniger Racks erreichen.3

Und

Ausgewählte AMD EPYC™ 9005 Prozessoren sind optimiert, um als Host-CPUs für GPU-fähige Systeme zu fungieren, wodurch die KI-Auslastungs-Performance gesteigert werden kann.4

Mehr entdecken



FÜHRENDE GENERATIVE KI-BESCHLEUNIGER AMD INSTINCTTM MI325X-SERIE

Setzt mit der AMD CDNA™ Architektur der 3. Generation neue Maßstäbe in der KI-Leistung und bietet unglaubliche Performance und Effizienz für Training und Inferenz.

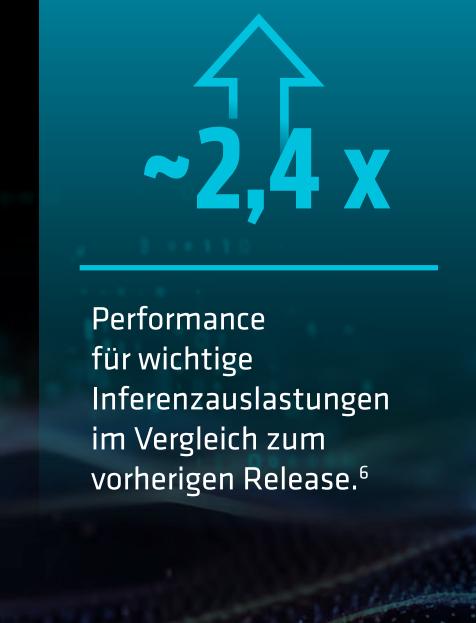
Speicher mit 256 GB und einer Bandbreite von bis zu 6 TB/s die Performance und trägt zur Senkung der Gesamtbetriebskosten bei. 5

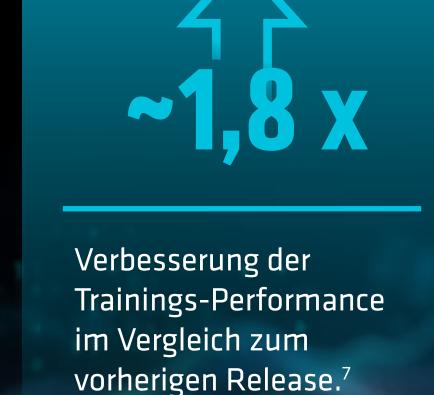
Optimiert mit einem branchenführenden HBM3E-

Mehr entdecken

SOFORT MAXIMALE PERFORMANCE AMD ROCM™ 6.2 SOFTWARE Offener Software-Stack, optimiert für generative KI- und HPC-Anwendungen, einschließlich Treibern,

Entwicklungstools und APIs für GPU-Programmierung – angefangen auf niedriger Ebene beim Kernel bis hin zu Anwendungen für Endbenutzer.







KI-NETWORKING MIT HOHER PERFORMANCE UND EINEM

Mehr entdecken

OFFENEN ÖKOSYSTEM Damit KI-Systeme funktionieren, sind zahlreiche Dienste erforderlich, die sich ständig weiterentwickeln, - vom Benutzerzugriff und der Datenaufnahme in die GPUs bis hin zur Einrichtung

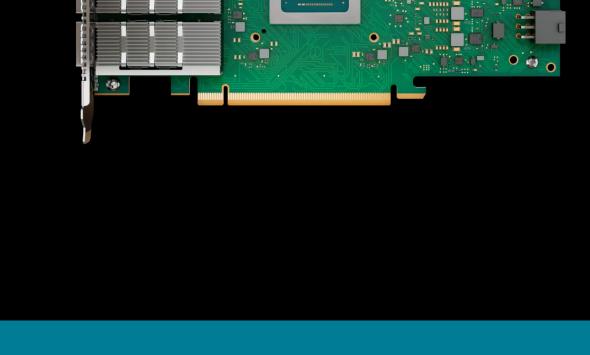
der KI-Engine, dem Start der Abfrage und der Bereitstellung der Ergebnisse. High-Performance-Networking stellt sicher, dass diese Funktionen flüssig ablaufen. AMD Pensando™ 400 Adapter Ermöglicht effiziente, zuverlässige und äußerst schnelle



gleichzeitig einen beachtlichen Durchsatz für P4-Pipelines.

Datenübertragungen, sodass Unternehmen das Potenzial des KI-Cloud-Computing voll ausschöpfen können.

Mehr entdecken



AMD RYZEN™ PRO PROZESSOREN FUR **NOTEBOOKS** Die weltweit erste dedizierte KI-Engine mit x86-Prozessoren treibt Hunderte unterschiedliche KI-Funktionen an und bringt professionelle Notebook-Benutzer

DIE POWER DER

KI AUF IHREM PC

dazu, Produktivität für ihre Workflows neu zu überdenken – all das mit unglaublicher Energieeffizienz und Geschwindigkeit. Mehr entdecken

Nvidia Blackwell-Spezifikationen unter https://resources.nvidia.com/en-us-blackwell-architecture?_gl=1*1r4pme7*_gcl_

Konfigurationen:

ABER EIN PORTFOLIO. Entdecken Sie die End-to-End-KI-Infrastrukturprodukte, -Lösungen und -Ökosysteme.

BEI KI-COMPUTING

EINHEITSLÖSUNG.

GIBT ES KEINE

Mehr erfahren

6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -I 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT = off, Determinism = Power, Turbo Boost = Enabled) 2P AMD EPYC 9654 (192 Kerne gesamt), 6 Instanzen mit 32 Kernen, NPS1, 1,5 TB 24 x 64 GB DDR5-4800, 1DPC, 2 x 1,92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT = off, Determinism = Power) Im Vergleich zu 2P Xeon Platinum 8592+ (128 Kerne gesamt), 4 Instanzen mit 32 Kernen, AMX Ein, 1 TB 16 x 64 GB DDR5-5600, 1DPC, 1,0 Gbit/s NetXtreme BCM5719 Gigabit Ethernet PCIe, 3,84 TB KIOXIA KCMYXRUG3T84 NVMe, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -I 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT = off, Determinism = Power, Turbo Boost = Enabled) CPU Medianwert Relativer Wert Generationenvergleich Turin 192 Kerne, 12 Instanzen 6067,531 3,775 2,278 Turin 128 Kerne, 8 Instanzen 4091,85 2,546 1,536 Genoa 96 Kerne, 6 Instanzen 2663,14 1,657 1

Die Ergebnisse können abhängig von Faktoren wie Systemkonfiguration, Softwareversion und BIOS-Einstellungen variieren. TPC, TPC Benchmark und TPC-C sind Marken des Transaction Processing Performance Council.

abgeleitet und als solcher nicht mit den veröffentlichten TPCx-AI-Ergebnissen vergleichbar, da die Ergebnisse des durchgängigen KI-Durchsatztests nicht der TPCx-AI-Spezifikation entsprechen.

6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT = off, Determinism = Power, Turbo Boost = Enabled)

2. 9xx5-040A: XGBoost (Läufe/Stunde) Durchsatzergebnisse basierend auf internen Tests von AMD vom 05.09.2024. XGBoost Konfigurationen: v2.2.1, Higgs Data Set, Instanzen mit 32 Kernen, FP32 2P AMD EPYC 9965 (384 Kerne gesamt), Instanzen mit 12 x 32 Kernen, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-45-generic (tuned-adm profile throughput-performance, ulimit -l 198078840, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT = off, Determinism = Power, Turbo Boost = Enabled), NPS = 12P AMD EPYC 9755 (256 Kerne gesamt), 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -I 198094956, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT = off, Determinism = Power, Turbo Boost = Enabled), NPS = 1 2P AMD EPYC 9654 (192 Kerne gesamt), 1,5 TB 24 x 64 GB DDR5-4800, 1 DPC, 2 x 1,92 TB Samsung MZQL21T9HCJR-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198120988, ulimit -n 1024, ulimit -s 8192), BIOS TTI100BA (SMT = off, Determinism = Power), NPS = 1 Im Vergleich zu 2P Xeon Platinum 8592+ (128 Kerne gesamt), AMX Ein, 1 TB 16 x 64 GB DDR5-5600, 1 DPC, 1,0 Gbit/s NetXtreme BCM5719 Gigabit Ethernet PCIe, 3,84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT = off, Determinism = Power, Turbo Boost = Enabled) Ergebnisse: CPU Lauf 1 Lauf 2 Lauf 3 Medianwert Relativer Wert Durchsatz Generationenvergleich 2P Turin 192 Kerne, NPS1 1565,217 1537,367 1553,957 1553,957 3 2,41 2P Turin 128C, NPS1 1103,448 1138,34 1111,969 2,147 1,725 2P Genoa 96C, NPS1 662,577 644,776 640,95 644,776 1,245 1 2P EMR 64C 517,986 421,053 553,846 517,986 1 NA Die Ergebnisse können abhängig von Faktoren wie Systemkonfiguration, Softwareversion und BIOS-Einstellungen variieren. 3. 9xx5TCO-001B: Dieses Szenario fußt auf vielen Annahmen und Schätzungen, und obwohl es auf internen Forschungen und bestmöglichen Näherungswerten von AMD basiert, dient es nur als Beispiel zur Veranschaulichung und sollte nicht als Grundlage für die

Entscheidungsfindung anstelle tatsächlicher Tests dienen. Das AMD Server and Greenhouse Gas Emissions TCO (Total Cost of Ownership) Estimator Tool – Version 1.12 – vergleicht die benötigten AMD EPYC™ und Intel® Xeon® CPU-basierten Server für insgesamt

1. 9xx5-012: TPCxAI @SF30 Multi-Instanz mit 32 Kernen Instanzgröße-Durchsatzergebnisse basierend auf internen Tests von AMD vom 05.09.2024 bei Ausführung mehrerer VM-Instanzen. Der aggregierte durchgängige KI-Durchsatztest ist vom TPCx-AI-Benchmark

2P AMD EPYC 9965 (384 Kerne gesamt), 12 Instanzen mit 32 Kernen, NPS1, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS,

2P AMD EPYC 9755 (256 Kerne gesamt), 8 Instanzen mit 32 Kernen, NPS1, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS,

39.100 Einheiten von SPECrate2017_int_base-Performance (Stand 10. Oktober 2024). Bei diesem Szenario wird ein älterer 2P-Server mit Intel Xeon Platinum_8280 mit 28 Kernen mit einer Bewertung von 391 verglichen mit einem 2P-Server mit EPYC 9965 (192 Kerne) und einer Bewertung von 3000 (https://www.spec.org/cpu2017/results/res2024q4/cpu2017-20240923-44837.pdf) sowie mit einem Vergleichs-Upgrade auf einen 2P-Server mit Intel Xeon Platinum 8592+ (64 Kerne) mit einer Bewertung von 1130 (https://spec.org/cpu2017/ results/res2024q3/cpu2017-20240701-43948.pdf). Die tatsächliche SPECrate®2017_int_base-Bewertung für 2P EPYC 9965 kann je nach OEM-Veröffentlichung abweichen. Schätzungen der Umweltauswirkungen auf der Grundlage dieser Daten unter Verwendung der länder- und regionenspezifischen Stromfaktoren aus "2024 International Country Specific Electricity Factors 10 – July 2024" und "Greenhouse Gas Equivalencies Calculator" der United States Environmental Protection Agency. 4. 9xx5-059A: Stable Diffusion XL v2 Trainingsergebnisse basierend auf internen Tests von AMD vom 10.10.2024. SDXL-Konfigurationen: DeepSpeed 0.14.0, TP8 Parallel, FP8, Batchgröße 24, Ergebnisse in Sekunden 2P AMD EPYC 9575F (128 Kerne gesamt) mit 8 x AMD Instinct MI300X-NPS1-SPX-192GB-750W, GPU Interconnectivity XGMI, ROCm™ 6.2.0-66, 2304 GB 24 x 96 GB DDR5-6000, BIOS 1.0 (Power Determinism = off), Ubuntu® 22.04.4 LTS, Kernel 5.15.0-72-generic,

2P Intel Xeon Platinum 8592+ (128 Kerne gesamt) mit 8 x AMD Instinct MI300X-NPS1-SPX-192GB-750, GPU Interconnectivity XGMI, ROCm 6.2.0-66, 2048 GB 32x64GB DDR5-4400, BIOS 2.0.4, (power determinism = off), Ubuntu 22.04.4 LTS, kernel 5.15.0-72-generic, 400,43 Sekunden für 19,600 % Steigerung der Trainings-Performance. Die Ergebnisse können abhängig von Faktoren wie Systemkonfiguration, Softwareversion und BIOS-Einstellungen variieren. 5. MI325-001A: Berechnungen durchgeführt vom AMD Leistungslabor am 26.09.2024 basierend auf aktuellen Spezifikationen und/oder Schätzungen. Der AMD Instinct™ MI325X OAM Beschleuniger wird über eine Speicherkapazität von 256 GB HBM3E und eine maximale theoretische GPU-Speicherbandbreite von 6 TB/s verfügen. Die tatsächlichen Ergebnisse auf Basis von Chips aus der Produktion können abweichen. Die höchsten veröffentlichten Ergebnisse für den NVidia Hopper H200 (141 GB) SXM-GPU-Beschleuniger ergaben eine Speicherkapazität von 141 GB HBM3e und eine GPU-Speicherbandbreite von 4,8 TB/s. https://nvdam.widen.net/s/nb5zzzsjdf/hpc-datasheet-sc23h200-datasheet-3002446

aw*R0NMLjE3MTM5NjQ3NTAuQ2p3S0NBancyNkt4QmhCREVpd0F1NktYdDlweXY1dlUtaHNKNmhPdHM4UVdPSIM3dFdQaE40Wkl4THZBaWFVajFy 6. MI300-062: Getestet am 29. September 2024 im AMD Leistungslabor zum Vergleich der Inferencing-Performance der ROCm 6.2 Software und der ROCm 6.0 Software auf Systemen mit acht AMD Instinct™ MI300X GPUs in Verbindung mit den Modellen Llama 3.1-8B, Llama 3.1-70B, Mixtral-8x7B, Mixtral-8x22B und Qwen 72B. Die Performance von ROCm 6.2 mit vLLM 0.5.5 wurde mit der Performance von ROCm 6.0 mit vLLM 0.3.3 verglichen, und es wurden Tests über Batchgrößen von 1 bis 256 und Sequenzlängen von 128 bis 2.048 durchgeführt.

1P AMD EPYC™ 9534 CPU Server mit 8 x AMD Instinct™ MI300X (192 GB, 750 W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA pro Sockel), 1,5 TiB (24 DIMMs, 4800 MT/s Speicher, 64 GiB/DIMM), 4 x 3,49 TB Micron 7450 Speicher, BIOS-Version: 1.8, ROCm 6.2.0-00, vLLM 0.5.5, PyTorch 2.4.0, Ubuntu[®] 22.04 LTS mit Linux-Kernel 5.15.0-119-generic. Im Vergleich zu

Die höchsten veröffentlichten Ergebnisse für den Nvidia Blackwell HGX B100 (192 GB) 700-W-GPU-Beschleuniger ergaben eine HBM3E-Speicherkapazität von 192 GB und eine GPU-Speicherbandbreiten-Performance von 8 TB/s.

Die höchsten veröffentlichten Ergebnisse für den Nvidia Blackwell HGX B200 (192 GB) GPU-Beschleuniger ergaben eine Speicherkapazität von 192 GB HBM3E und eine GPU-Speicherbandbreite von 8 TB/s.

Marken der Intel Corporation oder ihrer Tochtergesellschaften. Java® ist eine eingetragene Marke von Oracle und/oder seinen Partnern. SPEC®, SPEC CPU®, SPECrate®, SPECint® und SPECpower_ssj® sind eingetragene Marken der Standard Performance Evaluation Corporation. Weitere Informationen auf www.spec.org. Andere Produktnamen in diesem Dokument dienen nur zur Information und können Marken ihrer jeweiligen Inhaber sein. Bestimmte AMD Technologien erfordern möglicherweise die Ermöglichung oder Aktivierung durch Dritte. Die unterstützten Funktionen können je nach Betriebssystem

variieren. Bitte informieren Sie sich beim Systemhersteller über spezifische Funktionen. Keine Technologie und kein Produkt kann vollständig sicher sein.

1P AMD EPYC 9534 CPU Server mit 8 x AMD Instinct™ MI300X (192 GB, 750 W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA pro Sockel), 1,5 TiB 24 DIMMs, 4800 MT/s Speicher, 64 GiB/DIMM), 4 x 3,49 TB Micron 7450 Speicher, BIOS-Version: 1.8, ROCm 6.0.0-00, vLLM 0.3.3, PyTorch 2.1.1, Ubuntu 22.04 LTS mit Linux-Kernel 5.15.0-119-generic.

Serverhersteller wählen möglicherweise andere Konfigurationen, was zu anderen Ergebnissen führen kann. Die Performance kann je nach Faktoren wie u. a. verschiedenen Versionen von Konfigurationen, vLLM und Treibern variieren. 7. MI300-63: Tests durchgeführt im AMD Leistungslabor am 29. September 2024 zum Vergleich der Trainings-Performance der ROCm 6.2 Software und der ROCm 6.0 Software, beide mit Megatron-LM, auf Systemen mit 8 AMD Instinct™ MI300X GPUs mit Llama 2-7B, Llama 2-70B (4K), Qwen1.5-14B Modellen unter Verwendung benutzerdefinierter Docker-Container für jedes System. ROCm 6.2 mit Megatron-LM TFLOPs wurde verglichen mit den TFLOPs mit ROCm 6.0 mit Megatron-LM.

CPU: 1P AMD EPYC 9454 Prozessor mit 48 Kernen, Host-Speicher: 2 x 3,5 GB GPU: AMD Instinct MI300X 1P AMD EPYC™ 9454 CPU, 8 x AMD Instinct™ MI300X (192 GB, 750 W) GPUs, American Megatrends International LLC BIOS-Version: 1.8, ROCm 6.2 internes Release, Megatron-LM Codeverzweigungen hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_gwen1.5 for Qwen1.5-14B, PyTorch 2.4, Ubuntu 22.04 LTS mit Linux kernel 5.15.0-117-generic.

im Vergleich zu 1P AMD EPYC 9454 Prozessor mit 48 Kernen, 8 x AMD Instinct™ MI300X (192 GB, 750 W) GPUs, American Megatrends International LLC BIOS-Version: 1.8, ROCm 6.0.0, Megatron-LM Codeverzweigungen hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_qwen1.5 for Qwen1.5-14B, PyTorch 2.2, Ubuntu 22.04 LTS mit Linux kernel 5.15.0-117-generic.

Serverhersteller wählen möglicherweise andere Konfigurationen, was zu anderen Ergebnissen führen kann. Die Performance kann je nach Faktoren wie u. a. verschiedenen Versionen von Konfigurationen, Megatron-LM und Treibern variieren. Ergebnisse: MI300X mit ROCm 6.2 liefert im Durchschnitt das 1,83-Fache des Trainingsdurchsatzes (83 % höher) als ROCm 6.0. AMD © 2025 Advanced Micro Devices, Inc. Alle Rechte vorbehalten. AMD, das AMD Pfeillogo, EPYC und deren Kombinationen sind Marken von Advanced Micro Devices, Inc. Intel®, das Intel Logo und Xeon® sind

together we advance_