AMD
together we advance_

DAS KI-KONTINUUM:
WELCHE INFRASTRUKTUR EIGNET SICH AM BESTEN FÜR INFERENZ?

Die Einplanung von KI ist für jede Aktualisierung des Rechenzentrums unerlässlich. GPUs sind wichtig für große KI-Auslastungen, aber die neuesten Generationen von CPUs können neben allgemeinen Auslastungen eine Vielzahl von KI-Aufgaben übernehmen. Beachten Sie diese Aspekte, wenn Sie Ihre wachsenden KI-Inferenzanforderungen ermitteln.

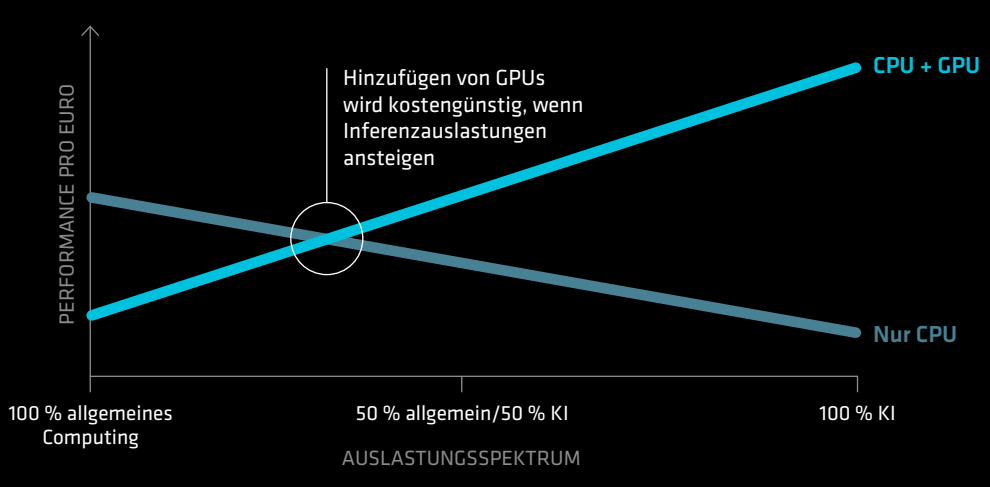
Oft benötigt KI keine Echtzeitergebnisse

Moderne CPUs können kleine bis mittlere KI-Inferenzauslastungen mit einer Latenz von weniger als einer Sekunde ausführen. Wenn KI-Inferenzauslastungen wachsen oder die Reaktionszeiten kürzer werden, müssen Sie möglicherweise diskrete Beschleuniger hinzufügen.

BATCH-VERARBEITUNG	MITTLERE LATENZ	NIEDRIGE LATENZ	NAHEZU ECHTZEIT	ECHTZEIT
Minuten bis Tage	Sekunden bis Minuten	~500 ms bis Sekunden	~100 ms bis ~500 ms	<10 ms bis ~50 ms
ANWENDUNGSFÄLLE				
• Dokumentverarbeitung und -klassifizierung • Data Mining und Analysen • Wissenschaftliche Simulationen	• Übersetzung • Indexierung • Content Moderation • Prädiktive Instandhaltung	• Virtuelle Assistenten • Chatbots • Expert Agents • Videountertitelung	• Betrugserkennung • Entscheidungsfindung • Dynamische Preisfindung • Audio- und Videofilterung	• Finanzhandel • Telekommunikation und Netzwerke • Autonome Systeme
CPUs	CPUs + GPUs	Mehrere GPU-Cluster		

Wenn KI-Auslastungen ansteigen, werden GPUs immer kostengünstiger

CPUs allein können gemischte Unternehmensauslastungen und KI unterstützen. Wenn Modellgröße, Komplexität und Volumen ansteigen, können GPU-Cluster mehr Performance pro Euro bieten.

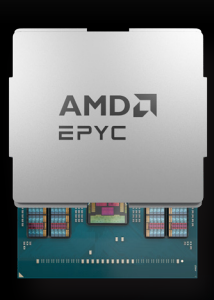


Die Grafik dient nur zu Illustrationszwecken. Der Schnittpunkt variiert je nach Auslastungen und Prozessormodellen.

Unterschiedliche Modelle haben einzigartige Verarbeitungsanforderungen

Maschinelles Lernen, Grafikverarbeitung und statistische Methoden werden auf CPUs außergewöhnlich gut ausgeführt. Kleine bis mittlere Large Language Models (LLMs) funktionieren gut auf den neuesten CPUs. Größere Modelle können einen deutlichen Nutzen aus KI-Beschleunigern ziehen.

	MODELLGRÖSSE	PROZESSOR	AMD LÖSUNG
Klassisches maschinelles Lernen	~1 MB bis ~200 MB	CPUs, integriert in das Rechenzentrum	AMD Ryzen™ CPUs AMD EPYC™ CPUs
Deep Learning	~60 Mio. Parameter bis ~20 Mrd.	CPUs, Rechenzentrum	AMD EPYC CPUs (hohe Kernzahl)
Transformer/LLMs	~20 Mrd. Parameter bis ~450 Mrd.	CPUs + GPUs	AMD EPYC CPUs (hohe Frequenz) + AMD Instinct™ GPUs **oder** NVIDIA GPUs
	~450 Mrd. Parameter und höher	Mehrere GPU-Cluster	



AMD EPYC CPUs brillieren mit KI für Unternehmen

AMD EPYC CPUs der 5. Generation bieten wichtige Performance-Verbesserungen für KI-Auslastungen:

Bis zu
3,8-facher
Durchsatz für End-to-End-KI im Vergleich zu CPUs der Konkurrenz[1]

Bis zu
90 %
schnellerer Durchsatz auf Llama 3.1 8B bei BF16 im Vergleich zu CPUs der Konkurrenz[2]

Bis zu
86 %
schnellere Facebook AI Similarity Search (FAISS) im Vergleich zu EPYC CPUs der vorherigen Generation[3]

AMD EPYC™ CPUs DER 5. GENERATION:
DIE BESTE CPU FÜR UNTERNEHMENS-KI[4]

Erfahren Sie, warum AMD EPYC CPUs der 5. Generation bei KI-Inferenzauslastungen brillieren.

EPYC für KI-Inferenz besuchen

1. 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 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. 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), 1 DPC, 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 -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT = off, Determinism = Power, Turbo Boost = Enabled) 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), 1 DPC, 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 -l 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, 1 DPC, 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, 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 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 EMR 64 Kerne, 4 Instanzen 1607,417 1 k. A. 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. (9xx5-012)
2. Llama3.1-8B-Durchsatzergebnisse basierend auf internen Tests von AMD vom 05.09.2024. Llama3-8B-Konfigurationen: IPEX.LLM 2.4.0, NPS = 2, BF16, Batch-Größe 4, Eingabe-/Ausgabe-Token-Konfigurationen (Anwendungsfälle): [Zusammenfassung = 1024/128, Chatbot = 128/128, Übersetzung = 1024/1024, Essay = 128/1024, Beschriftung = 16/16]. 2P AMD EPYC 9965 (384 Kerne gesamt), 6 Instanzen mit 64 Kernen, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1 DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.3 LTS, 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), NPS = 2 2P AMD EPYC 9755 (256 Kerne gesamt), 4 Instanzen mit 64 Kernen, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1 DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.3 LTS, 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), NPS = 2 2P AMD EPYC 9654 (192 Kerne gesamt) 4 Instanzen mit 48 Kernen, 1,5 TB 24 x 64 GB DDR5-4800, 1 DPC, 1,0 Gbit/s NetXtreme BCM5720 Gigabit Ethernet PCIe, 3,5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 5.15.85-051585-generic (tuned-adm profile throughput-performance, ulimit -l 1198117616, ulimit -n 500000, ulimit -s 8192), BIOS RVI1008C (SMT = off, Determinism = Power, Turbo Boost = Enabled), NPS = 2 im Vergleich zu 2P Xeon Platinum 8592+ (128 Kerne gesamt), 2 Instanzen mit 64 Kernen, 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 2P EMR 64 Kerne 2P Turin 192 Kerne 2P Turin 128 Kerne 2P Genoa 96 Kerne Durchschnitt Aggregiert Medianwert Gesamtdurchsatz 99,474 193,267 182,595 138,978 Konkurrenz 1 1,943 1,836 1.397 Generationenvergleich k. A. 1,391 1,314 1. Die Ergebnisse können abhängig von Faktoren wie Systemkonfiguration, Softwareversion und BIOS-Einstellungen variieren. (9xx5-009)
3. FAISS (Anfragen/Stunde) Durchsatzergebnisse basierend auf internen Tests von AMD vom 05.09.2024. FAISS-Konfigurationen: sift1m-Datensatz, Instanzen mit 16 Kernen, FP32, MKL 2024.2.1 2P AMD EPYC 9965 (384 Kerne gesamt), 24 Instanzen mit 16 Kernen, 1,5 TB 24 x 64 GB DDR5-6400 (bei 6000 MT/s), 1 DPC, 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 -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT = off, Determinism = Power, Turbo Boost = Enabled), NPS = 4 2P AMD EPYC 9654 (192 Kerne gesamt) 12 Instanzen mit 16 Kernen, 1,5 TB 24 x 64 GB DDR5-4800, 1 DPC, 2 x 1,92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT = off, Determinism = Power), NPS = 4 im Vergleich zu 2P Xeon Platinum 8592+ (128 Kerne gesamt), 8 Instanzen mit 16 Kernen, 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 Medianwert Relativer Wert Durchsatz Generationenvergleich 2P Turin 192 Kerne 64,2 3,776 1,861 2P Genoa 96 Kerne 34,5 2,029 1 2P EMR 64 Kerne 17 1 k. A. Die Ergebnisse können abhängig von Faktoren wie Systemkonfiguration, Softwareversion und BIOS-Einstellungen variieren. (9xx5-011)
4. Vergleich basiert auf Thread-Dichte, Performance, Funktionen, Prozesstechnologie und integrierten Sicherheitsfunktionen der derzeit bereitgestellten Server mit Stand vom 10.10.2024. EPYC 9005-Serie CPUs bieten die höchste Thread-Dichte, sind mit über 500 Weltrekorden branchenführend, einschließlich Weltrekord-Unternehmensführung bei Java® ops/sec-Performance, Top-HPC-Führung mit Gleitkommadurchsatz-Performance, KI-End-to-End-Performance mit TPCx-AI-Performance und höchster Energieeffizienzwerte. Im Vergleich zu Xeon der 5. Generation hat die EPYC-Serie der 5. Generation auch mehr DDR5-Speicherkanäle mit mehr Speicherbandbreite, unterstützt mehr PCIe® Gen5-Lanes für E/A-Durchsatz und hat bis zu 5-mal mehr L3-Cache/Kern für einen schnelleren Datenzugriff. Die EPYC 9005-Serie verwendet die erweiterte Technologie mit 3-4 nm und bietet die Sicherheitsfunktionen für sichere Speicherverschlüsselung, sichere verschlüsselte Virtualisierung(SVV), SVV Encrypted State und SVV Secure Nested Paging. (EPYC-029D)

AMD
together we advance_