SERVERHERO

WHITE PAPER

HPC MIT GPU-COMPUTING

Auf dem Weg ins Exascale-Zeitalter mit GPU-Computing

Dank immer schnellerer Computertechnik wird die Berechnung von äußerst großen Datenmengen im immer kürzerer Zeit möglich. So werden umfangreiche wissenschaftliche Aufgabenstellungen durch Simulationen, Modelle und Analysen lösbar, die vorher nicht umsetzbar waren.
Dies betrifft medizinische Forschung, Materialforschung, Meteorologie und die Auswertung von Big Data zu Smart Data sowie Anwendungen mit Künstlicher Intelligenz (KI), wie Maschinelles Lernen (ML) und Deep Learning (DL). All dies wird unter dem Begriff „High Performance Computing (HPC)“ zusammengefasst. Hierfür werden zuverlässige und sichere Server mit guter Leistung eingesetzt, wie zum Beispiel die ProLiant-Server von HPE mit hohen Speichergeschwindigkeiten von bis zu DDR4 3200 MT/s und bis zu 4 TB pro CPU.

Für höchst umfangreiche Analysen bedarf es Supercomputer. Supercomputing-Cluster können Zehntausende Prozessoren beinhalten. Doch reicht deren Leistung bei hochkomplexen Berechnungen allein nicht aus, daher verdichten die meisten HPC-Systeme zur parallelen Verarbeitung mehrere Prozessoren und Speichermodule durch ultrahohe Bandbreitenverbindungen. Bei einigen HPC-Systemen werden CPUs und GPUs zusammen eingesetzt (heterogenes Rechnen)[1]. Dabei werden die rechenintensivsten Bereiche der Anwendung auf dem Grafikprozessor ausgeführt, die restlichen, weniger anspruchsvollen Aufgaben erledigt die CPU.
Anders als CPUs können GPUs über tausende Cores für die parallele Datenverarbeitung haben. Zusammen mit der seriellen Verarbeitung der CPU-Aufgaben können Anwendungen viel schneller umgesetzt werden.

So funktioniert HPC

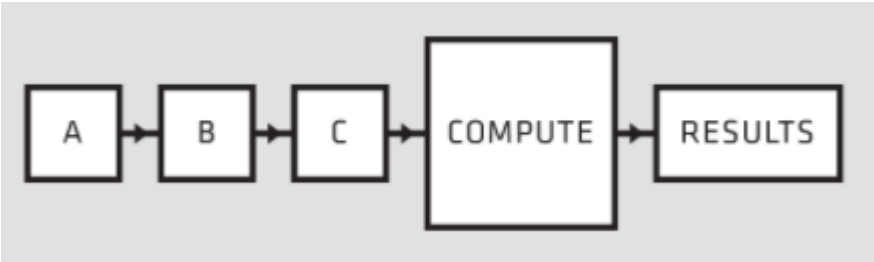
Die beiden hauptsächlichen Methoden zur Informationsverarbeitung des HPC sind serielle und parallele Verarbeitung.

Serielle Verarbeitung

Jeder CPU-Kern führt in der Regel nur jeweils eine Aufgabe aus.

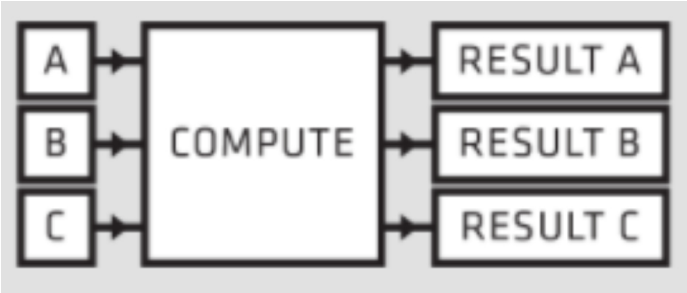
[1] Heterogenes Rechnen: HPC-Architektur, die serielle (CPU) und parallele (GPU) Verarbeitungskapazität optimiert.

CPUs sind für Verarbeitungsfunktionen wie Betriebssysteme und wichtigste Anwendungen (z. B. Textverarbeitung, Bürosysteme) unverzichtbar.



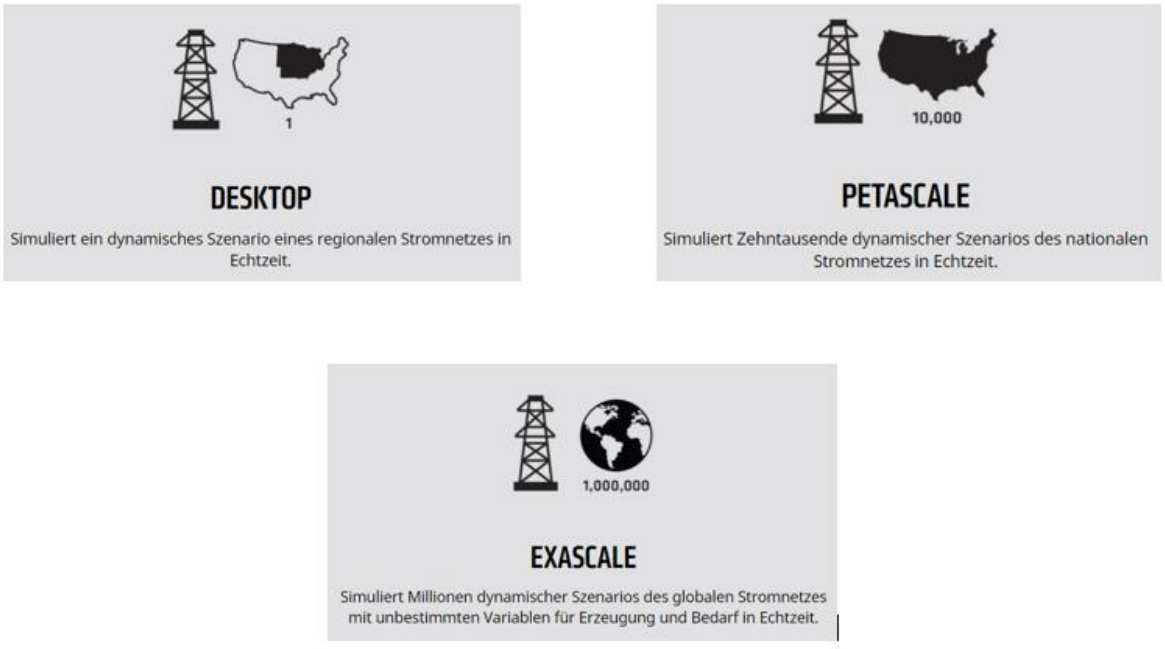
Serielle Datenverarbeitung (AMD)

Parallele Verarbeitung
Dies wird durch Verwendung mehrerer CPUs oder Grafikprozessoren (GPUs) möglich. Die ursprünglich für Grafikkarten konzipierten GPUs können mehrere arithmetische Operationen über eine Datenmatrix (z. B. Bildschirmpixel) gleichzeitig durchführen. Die Möglichkeit, auf zahlreichen Datenebenen gleichzeitig zu arbeiten, prädestiniert GPUs zur parallelen Verarbeitung bei Aufgaben für maschinelles Lernen, wie zum Beispiel das Erkennen von Objekten in Videos.



Parallele Datenverarbeitung (AMD)

Die Rechenleistung von Computern wird u. a. in FLOPS (*Floating Point Operation per Second*, dt. Gleitkommaoperationen pro Sekunde) gemessen. Anfang 2019 lag die Leistung eines Hochleistungs-Supercomputer bei 143,5 PetaFLOPS. Dies entspricht der Petascale-Klasse von Supercomputern, die über eine Billiarde FLOPS erreichen. Zum Vergleich: Ein Hochleistungs-Gaming-PC mit 200 GigaFLOPS ist etwa 1.000.000 mal langsamer.
Doch wird hinsichtlich Rechenleistung und -geschwindigkeit der Schritt zur Exascale-Marke mit 1.000 Mal schnelleren Berechnungen als bisher (Petascale) unternommen. Exascale-Supercomputer können dann 10^{18} Bytes (oder 1 Mrd. x 1 Mrd.) Berechnungen pro Sekunde durchführen.



Rechenleistungen (AMD)

Anwendungen für HPC

Durch die kontinuierliche Weiterentwicklung der Computertechnologien können heute wesentlich mehr und komplexere Fragestellungen bearbeitet werden. Dazu gehören unter anderem:

Maschinelles Lernen:

Maschinelles Lernen (ML) ist ein Bereich der der künstlichen Intelligenz (KI) und bezeichnet ein System, das selbstständig lernen kann und nicht nur passiv eingegebene Befehle ausführt. HPC-Systeme können mittels hochentwickeltem ML große Datenmengen analysieren, um z. B. in der Krebsforschung Melanome auf Fotos zu erfassen (Bildanalyse und -überwachung), erste echte Anzeichen von Problemen in Teilen oder in Maschinen zu erkennen (prognostizierte Instandhaltung), und Texte richtig zu übersetzen (Textanalyse und -klassifizierung), etc.

Big-Data-Analyse:

Riesige Datenmengen müssen in Wissenschaft, Finanzwesen, Wirtschaft und im Gesundheitssektor, aber auch bei Fragen der Netz- und Computersicherheit sowie in Behörden auf nationaler und internationaler Ebene schnell berechnet, verglichen und zusammengeführt werden. Dies erfordert entsprechend viele Durchsatz- und Rechenkapazitäten.

Beispiel:

Die NASA hat ein jährliches Volumen von schätzungsweise 50 Petabytes Missionsdaten und kann diese Datenmengen nur mit Supercomputing analysieren, um Modelle und Simulationen zu erstellen.

Erweiterte Modellerstellung und Simulation:

Unternehmen können auf kostenintensive Prototypenkonstruktionen verzichten, da die Anforderungen in realitätsgetreuen Modellen durchsimuliert werden können. So können Zeit, Material und Kosten gespart sowie neue Produkte schneller auf den Markt eingeführt werden.
Weitere Einsatzgebiete für HPC sind Energiesektor, Arzneimittelforschung, Automobilbranche, Luft- und Raumfahrt sowie die Berechnung von Klima- und Wettersystemen.

GPUs und CPUs in Supercomputern

Einen Meilenstein in der Entwicklung besserer Technologien für High Performance Computing brachte der Hersteller AMD, dessen GPUs (AMD Instinct MI200-Beschleuniger) sowie die Ankündigung einer neuen Generation CPUs (AMD EPYC „Genoa“ mit ZEN-4-Architektur im 5-nm-Verfahren) für Furore sorgten, da die Produkte in der Leistung weit vor allen anderen vergleichbaren CPUs und GPUs liegen.

Neue Spitzenleistungen mit GPU-Beschleuniger Instinct MI200 (AMD)

Aufgrund dieser technologischen Überlegenheit wird auch klar, warum Wissenschaftler bei der Entwicklung von Supercomputern auf AMD setzen, wie die folgenden Beispiele zeigen:

1. Frontier

AMD hat in Zusammenarbeit mit dem US-Energieministerium, dem Oak Ridge National Laboratory und Cray Inc. den Supercomputer Frontier entwickelt, der eine Spitzenrechenleistung von mehr als 1,5 ExaFLOPS liefern soll. Die Inbetriebnahme ist für 2022 vorgesehen.

Supercomputer Frontier (AMD)

Frontier ist konzipiert mit:

- für HPC und KI optimierte CPUs (AMD EPYC Gen3)
- speziell entwickelte High Bandwidth Memory-(HBM)-fähige GPUs (ADM Instinct MI250X)
- CPU-CPU-Verbindung mit AMD Infinity Fabric
- Netzwerkverbindung mit mehreren Slingshot-NICs; dies ermöglicht eine Netzwerkbandbreite von 100 GB/s. Das Slingshot-Netzwerk bietet adaptives Routing, Überlastungsmanagement und Dienstqualität (Quality of Service, QoS).

Frontier wird die Grenzen wissenschaftlicher Entdeckungen verschieben und die Reichweite der Forschung erweitern, indem der Supercomputer die Leistung von künstlicher Intelligenz (KI), Analysen und Simulationen im großen Maßstab steigert und Wissenschaftlern dabei hilft, mehr Berechnungen durchzuführen, neue Muster in Daten zu erkennen und innovative Datenanalysemethoden zu entwickeln, um wissenschaftliche Entdeckungen zu beschleunigen.

2. El Capitan

Das US-Energieministerium, das Lawrence Livermore National Laboratory und HPE haben sich mit AMD zusammengetan, um El Capitan zu entwickeln, den schnellsten Supercomputer der Welt, der Anfang 2023 ausgeliefert werden soll, der mehr als 2 ExaFLOPS mit doppelter Genauigkeit[2] erreichen soll. Beim El Capitan werden hochmoderne Produkte verbaut, in die Verbesserungen aus dem kundenspezifischen Prozessordesign von Frontier eingeflossen sind:

Supercomputer El Capitan (AMD)

- AMD EPYC Prozessoren der nächsten Generation (Codename „Genoa“) mit „Zen 4“-Prozessorkernarchitektur, um Speicher- und I/O-Subsysteme der nächsten Generation für KI- und HPC-Workloads zu unterstützen.

- Die AMD GPUs der nächsten Generation basieren auf einer neuen rechenoptimierten Architektur für HPC- und KI-Workloads und nutzen Speicher der nächsten Generation mit hoher Bandbreite für optimale Deep-Learning-Leistung

Dieses neue Design ist für die Analyse von KI- und ML-Daten wichtig, um Modelle zu erstellen, die schneller und genauer sind und die Unsicherheit ihrer Vorhersagen quantifizieren können.

---

[2] Doppelte Genauigkeit (englisch *double precision*) steht in der Computerarithmetik für ein Gleitkomma-Zahlenformat, bei dem eine Zahl 8 Byte (also 64 Bit) belegt. Es belegt damit doppelt so viel Speicher wie Gleitkommazahlen einfacher Genauigkeit.

„Wenn alle 7,7 Milliarden Menschen auf der Erde jeweils eine Berechnung pro Sekunde ausführten, würde es mehr als 8 Jahre dauern, um das zu erreichen, was El Capitan in einer Sekunde schafft.“

Hewlett Packard Enterprise zum Supercomputer El Capitan

Innovatives Design senkt Kosten für HPC

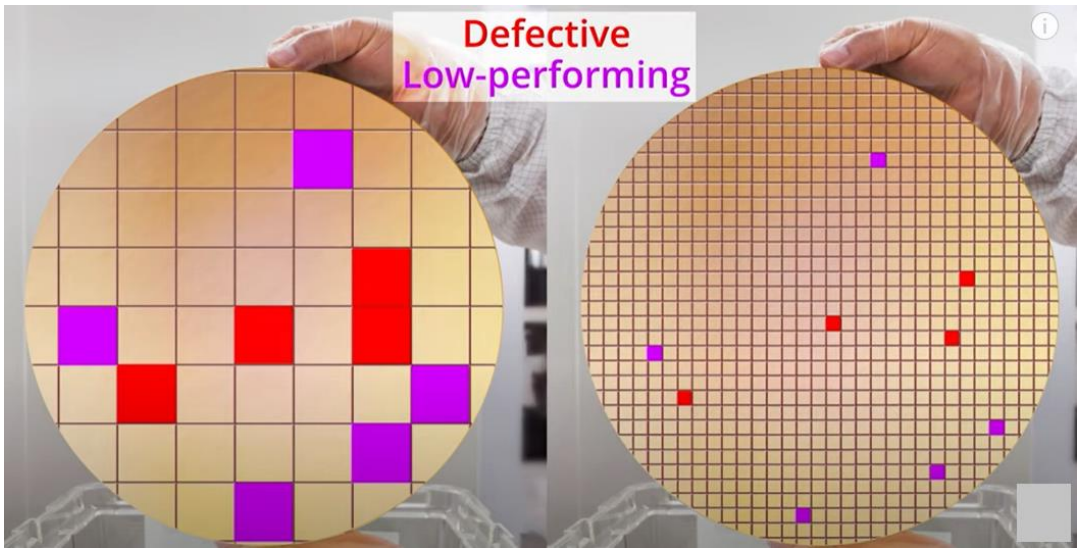
Wie ist es AMD nach Jahren der Vorherrschaft von Intel gelungen, sich an die Spitze der Branche im Bereich CPU und GPU zu setzen, sodass der Hersteller interessant für HPC-Anwendungen wurde?

CPU

AMD hat sich vom ausschließlich monolithischen One-*Die*-Design der Prozessoren abgewandt und setzt auf Chiplets. Dies sind modulare Mini-*Dies*, die zusammen in einer CPU untergebracht werden. Die CPU wird dadurch zwar größer, dies ist jedoch angesichts der erheblichen Leistungssteigerungen vernachlässigbar.

Die Vorteil von kleineren Chiplets liege in der Herstellung selbst: Die für Prozessoren benötigten Halbleitermaterialien werden in Form von sog. Silizium-Wafers gefertigt. Dabei kommt es zu Defekten und/oder suboptimal ausgeführten Bereichen auf den Wafern, sodass nicht das gesamte Material verwendbar ist.

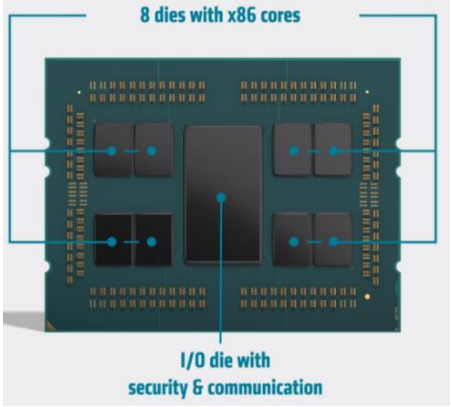
Durch die geringere Dimensionierung der Chiplets kann aus den Wafern mehr Material „herausgeholt“ und die Produktion beschleunigt werden. Dies verdeutlicht diese Abbildung:



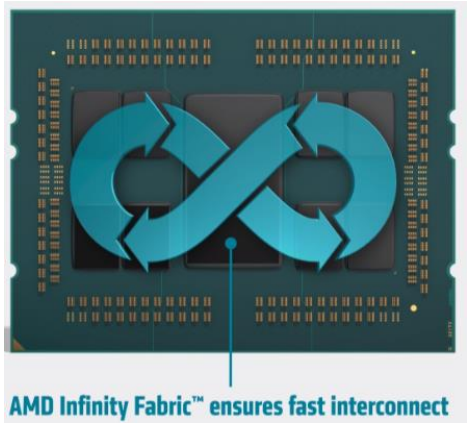
Mehr Ausbeute mit kleinerem Chiplet-Design (TechQuickie)

Allein dadurch ergeben sich für AMD erhebliche Einsparungen in der Produktion ihrer CPUs.

Weitere Vorteile liegen in der Chiplet-Technologie selbst: Durch mehrere kleine *Dies* (die wiederum mehrere Kerne haben) statt eines großen Mono-*Dies* können bestimmte Bereiche bzw. Aufgaben, die nicht reine Zahlen-berechnungen sind, aus der CPU ausgelagert werden, wie z. B. I/O und machen sie verfügbar für andere Aufgaben. Der Signalaustausch erfolgt über die AMD Infinity Fabric.



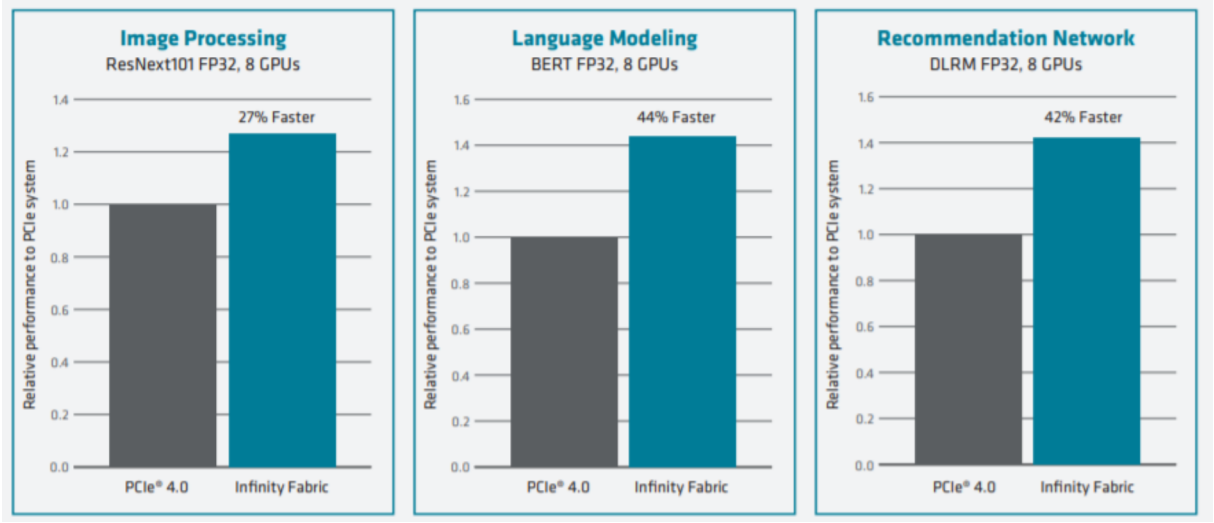
Chiplet-Technologie (AMD)



AMD Infinity Fabric (AMD)

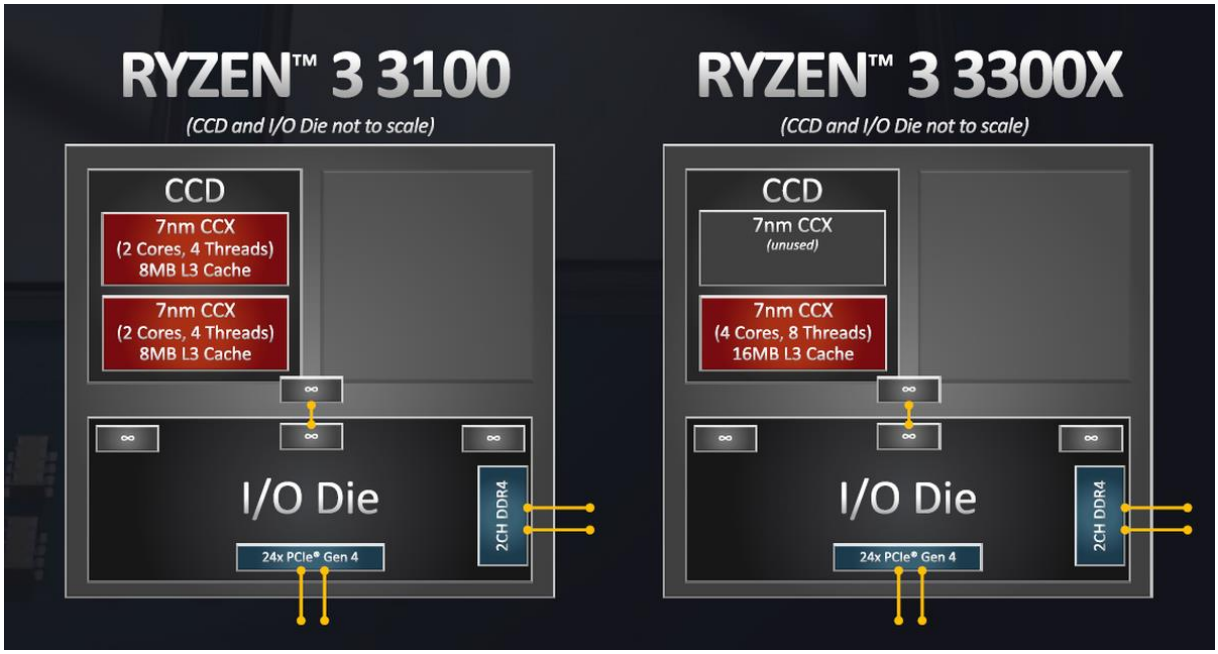
Zwar müssen die Signale der einzelnen *Dies* räumliche Abstände überwinden, was zu Verzögerungen führen kann, jedoch hat AMD mit seiner Infinity Fabric auch hier eine schnellere Lösung zur Hand als PCIe 4.0. Diese hybride Multi-*Die*-Architektur entkoppelt zwei Streams: acht *Dies* für die Prozessorkerne und ein I/O-*Die*, der für die prozessor-externe Sicherheit und Kommunikation zuständig ist. Das macht eine agile Bereitstellung einer neuartigen Prozess-technologie für CPU-Kerne möglich. Unabhängig davon können sich die I/O-Schaltkreise in ihrem eigenem Tempo entwickeln.

Neben einer leistungsstärkeren CPU wurde hier also auch die Geschwindigkeit der Signalübertragung gegenüber dem Standard PCIe 4.0 verbessert.



Benchmarks zu Signalübertragungsgeschwindigkeit: PCIe 4.0 vs. AMD Infinity Fabric (AMD)

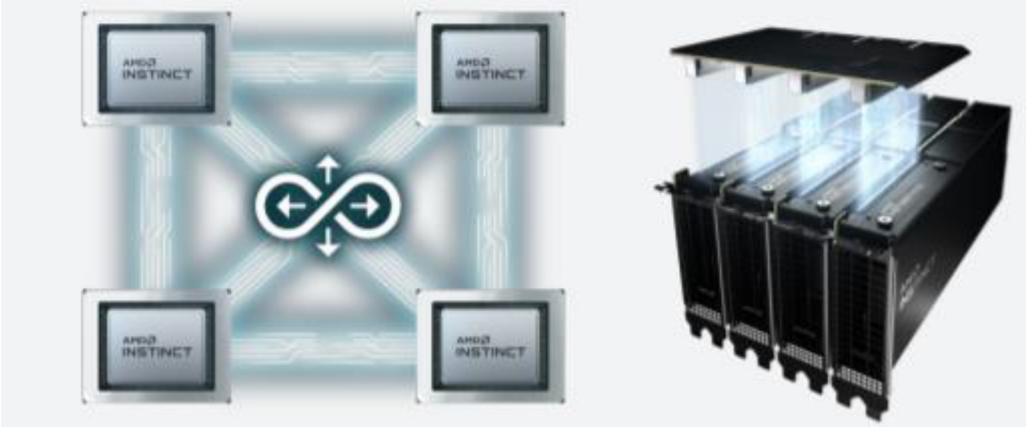
So ergeben sich mit der neuen Architektur Untergruppierungen, wie CCD (Core-Chiplet-*Dies*), CCX (CPU-Core-Complex), IOD (Input-Output-*Dies*) bzw. cIOD (Client IO-*Dies*), hier gezeigt am Beispiel einer CPU der Reihe AMD Ryzen 3 3100 und AMD Ryzen 33000X:



Chiplet-Architektur: AMD Ryzen 3 3100 und AMD Ryzen 3 3000X (AMD)

GPU

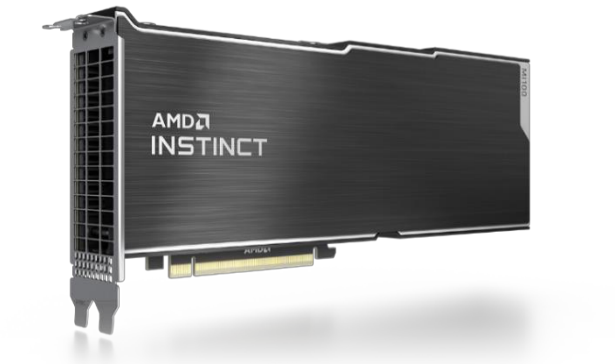
AMD hat eigens für HPC-Anwendungen einen speziellen Grafikprozessor bzw. -beschleuniger entwickelt: AMD Instinct MI200.
Als Nachfolger der ohnehin schon hochperformanten MI100-GPU steigert sie die Rechenleistung für hochkomplexe Analysen und Berechnungen in Wissenschaft und Forschung auf ein neues Niveau im Exascale-Bereich, um bessere Prognosen und Maßnahmen für den Klimawandel zu ermöglichen sowie die medizinische Forschung, speziell bei der Corona-Impfstoffentwicklung, voranzutreiben.
Dafür wurde sie mit der CDNA2-Architektur konfiguriert, d. h. diese GPU ist nicht für den Alltag oder für Gaming gedacht, sondern für rechenintensive Anwendungen im Exascale-Bereich. Laut AMD ist es „die weltweit schnellste HPC-GPU".



M200 Instinct GPUs mit Infinity Fabric (AMD)

Die MI100 bot als erster x86-Server-Grafikprozessor mit bis zu 46,1 TeraFLOPS einfacher Genauigkeit (FP32) schon Spitzenleistung für KI- und ML-Workloads. Im Vergleich dazu liefert die Nachfolgerin MI200 jedoch bis zu 95,7 TeraFLOPS doppelter Genauigkeit (FP64)!

In Kombination mit der Infinity Fabric und der ROCm-Open-Platform-Software, die offene Computersprachen, Compiler, Bibliotheken und Tools für die HPC-Community bietet, setzt die GPU neue Maßstäbe für HPC-Anwendungen wie Künstliche Intelligenz (KI), Maschinelles Lernen (ML), Deep Learning, etc.



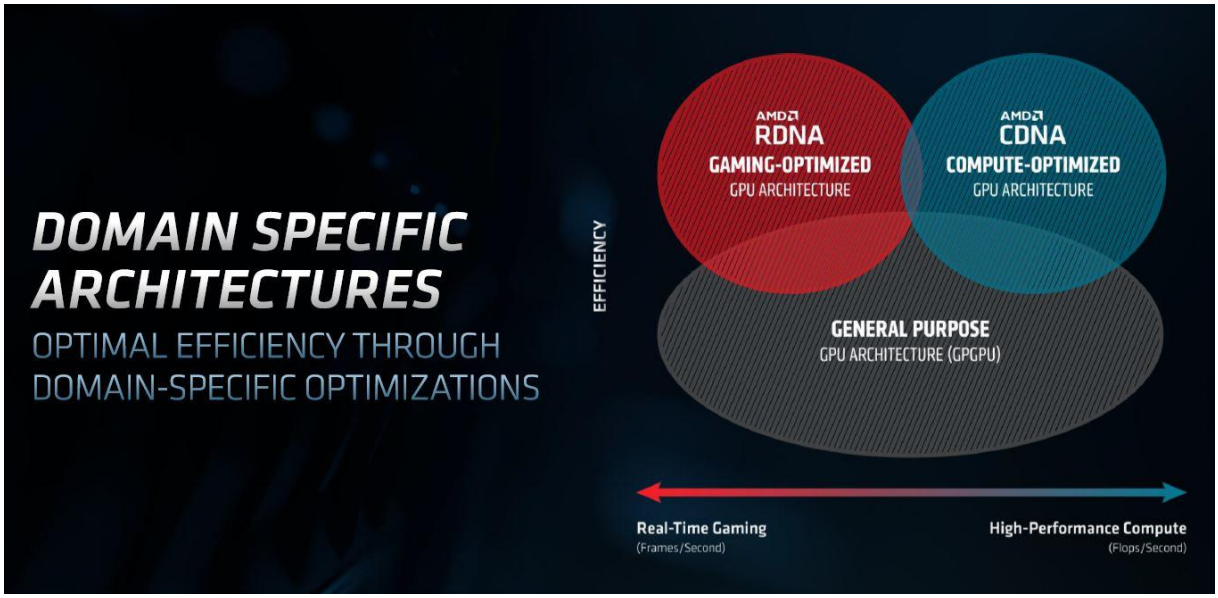
AMD Instinct GPU (AMD)

Bei der Entwicklung der AMD Instinct Beschleuniger lag der Fokus auf den Einsatz im wissenschaftlichen Bereich in Rechenzentren, um mit der schnelleren Verarbeitung von anspruchsvollen Workloads bei KI und Hochleistungsrechnen (High Performance Computing, HPC) neue Erkenntnisse und Lösungen voranzubringen. Die AMD Instinct Beschleuniger bieten so viel Leistung, dass sie nicht nur in herkömmlichen Anwendungen als Einzelserver, sondern auch in den größten Supercomputern der Welt verwendet werden. Mit Verbesserungen und Innovationen, wie AMD CDNA 2-Architektur und AMD Infinity Fabric-Technologie sind die neuesten AMD Instinct-Beschleuniger für Entdeckungen im Exascale-Bereich ausgelegt, um der Wissenschaft die Datengrundlage zu bieten, um vom Klimawandel bis zur Impfstoffforschung die größten Problem unserer Zeit anzugehen.

So liefert die AMD Instinct MI250X mit bis zu 47,9 TFLOPs doppelter Genauigkeit (FP64) und mit der neuen FP64-Matrix-Core-Technologie bis zu 95,7 TFLOPs doppelter Genauigkeit (FP64-Matrix) die vierfache Leistung im Vergleich zu anderen GPUs auf dem Markt. Für sehr komplexe KI-Workloads bei Maschinellem Lernen und Deep Learning erreicht die MI250X eine theoretische Spitzenleistung von bis zu 383 TFLOPS halber Genauigkeit (FP16) mit bis zu 1,6-mal mehr Speicherkapazität und Bandbreite als andere Grafikkarten.
AMD Instinct Beschleuniger mit neuer AMD CDNA2 Architektur sorgen dank Matrix Core Technologie mit 880 Matrix Cores in MI200 OAM-Beschleunigern für gesteigerte Rechenkapazität sowie für Unterstützung von mehr Anwendungen und Datentypen. Der deutlich höhere Datendurchsatz wurde ermöglicht, indem die AMD Instinct GPU als Multi-Chip-GPU konstruiert wurde. Zusätzlich kann diese hohe Leistung schnell von der CPU genutzt werden, da CPU und GPU für Cache-Kohärenz direkt miteinander verbunden sind.

Die CDNA2-Architektur von AMD

Grundsätzlich verfolgt AMD zwei Wege bei seiner GPU-Architektur:
Gaming-optimierte GPUs (RDNA) und Rechenlast-optimierte GPUs (CDNA), die beide auf der allgemeinen General-Purpose-GPU-Architektur (GPGPU) aufbauen.

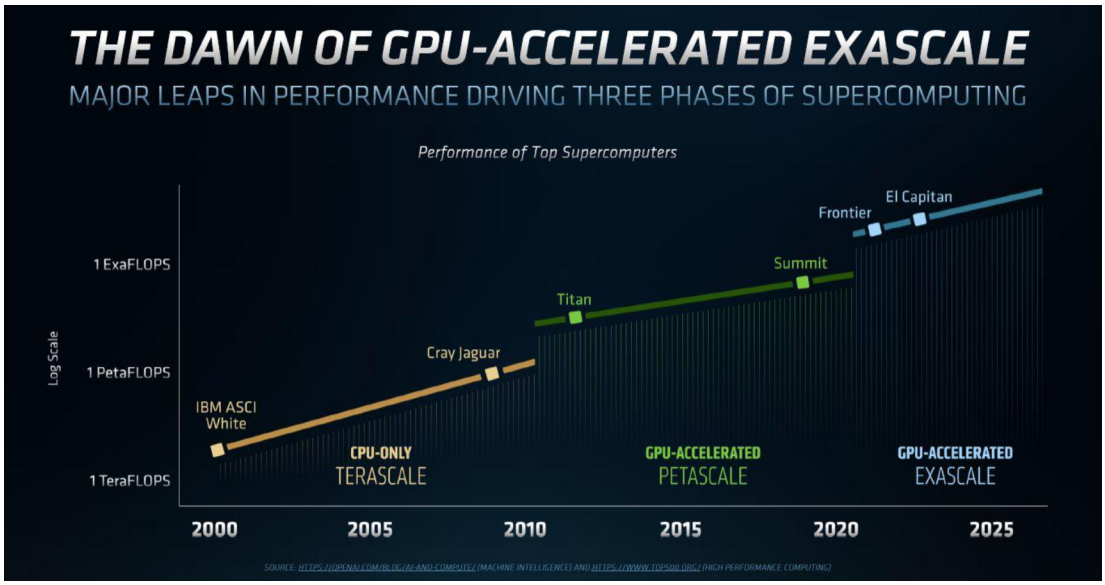


CPU-Arten bei AMD: RDNA, CDNA, GP (AMD)

Die CDNA2-Architektur wurde mit Rechnerkernen für das Zeitalter der Hochgeschwindigkeitsrechner entwickelt. Frühere Beschleunigerarchitekturen haben Leistung und Effizienz kontinuierlich verbessert und wurden gleichzeitig immer besser programmierbar. Die AMD CDNA2-Architektur hebt diesen Evolutionspfad auf die nächste Stufe und erreicht eine Leistungssteigerung um mehr als das Vierfache gegenüber der vorherigen Generation der AMD CDNA-Architektur, mit einem FP64-Doppelpräzisions-Vektordurchsatz von 47,9 TFLOP/s in der Spitze, um Exascale-Leistungsniveaus mit konkurrenzloser Programmierbarkeit in heterogenen Systemen zu ermöglichen.

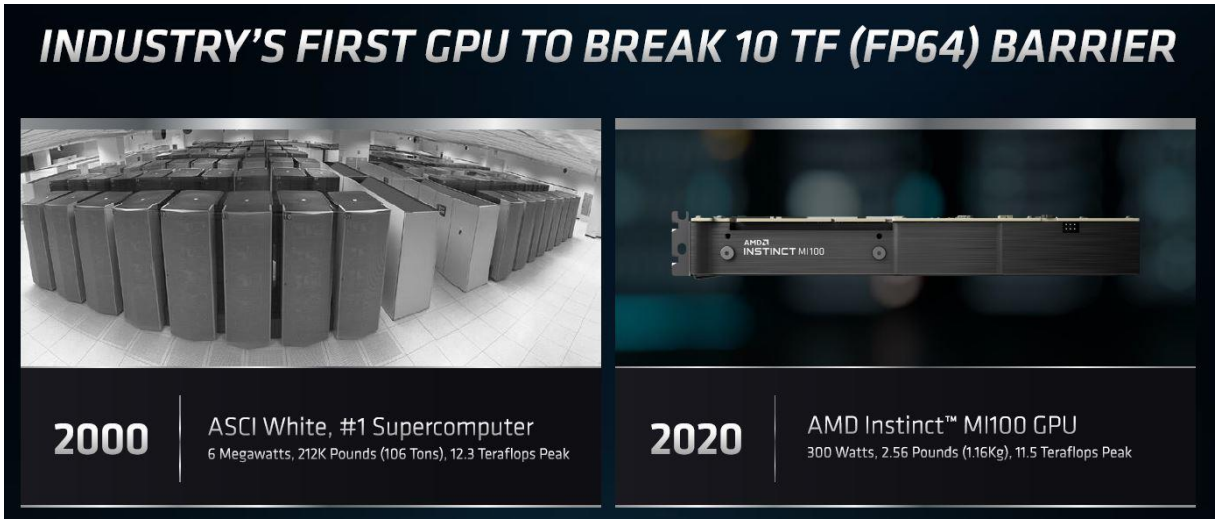
Hochleistungsrechnen mit CPU und GPU im Exascale-Zeitalter

Der gestiegene Bedarf an Rechenleistung in Wissenschaft, Finanzen und Industrie und die damit anfallenden riesigen Datenmengen verlangen nach entsprechenden Hardware-Lösungen, um den Herausforderungen der Zukunft, wie High Performance Computing (HPC), darunter ML, DL, KI, etc., begegnen zu können.



Entwicklung der Rechenleistung von Supercomputern (AMD)

Eine heutige moderne GPU kann einen Supercomputer aus dem Jahr 2000 hinsichtlich der Rechenleistung nahezu vollständig ersetzen, wie die folgende Abbildung verdeutlicht:



Leistungsvergleich Supercomputer vs. moderne GPU, hier noch mit Vorgänger-GPU AMD Instinct MI100 (AMD)

Die Branche ist also auf einem guten Weg, um die künftigen Rechenleistungsbedarfe auf dem Weg in das Exascale-Zeitalter zu erfüllen. Aktuell ist der Hersteller AMD mit seinen aufeinander abgestimmten Produkten (CPUs und GPUs) und der erforderlichen, leistungsstarken Umgebung (offene Software-Plattform ROCm, CDNA2-Architektur mit Infinity Fabric) in diesem Bereich für Nutzer, wie Wissenschaftseinrichtungen, Finanzwelt und Industrie mit Abstand am attraktivsten, da AMD seine qualitativ hochwertigen Produkte auch zu einem aktuell unschlagbaren Pries anbietet. Dies erleichtert Forschungsprogramme, wie zum Beispiel am Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee (USA):

- Berechnungen und Simulationen in der Medizin, speziell in der wichtigen Impfstoffforschung gegen Covid-19
- Simulationen in der Plasmaphysik für die nächste Generation von Anwendungen in der Strahlentherapie gegen Krebs sowie für die Erforschung der Molekularstruktur mittels Röntgenstrahlen in den Material- und Lebenswissenschaften
- CHOLLA-Projekt: Weiterentwicklung von Galaxis-Simulationen in der Astrophysik und der Erforschung des Ursprungs und des Entstehens von Galaxien
- u.v.m.

Darüber hinaus wird es in diesem Bereich interessante (Weiter-)Entwicklungen geben, die sich auch auf diese Gebiete auswirken werden:

- Cloud / Hyperscale
- Finanzen
- Energie
- Verstärkungslernen (Reinforcement Learning)
- Gesundheitswesen (Life Sciences)
- Automotive, Luft- und Raumfahrt
- HPC
- Bild- und Videoerkennung und -klassifizierung

Die (nahe) Zukunft bleibt also spannend!

Wenn Sie Fragen zu High Performance Computing haben, können Sie sich jederzeit an unsere zertifizierten Experten wenden, die Sie per Mail, Telefon oder im Chat herstellerunabhängig beraten!