BIT/ capital.

BIT Capital Research

Das Zeitalter der Künstlichen Intelligenz

September 2023

BIT capital RESEARCH

BIT CAPITAL RESEARCH WHITEPAPER

Infrastruktur des KI-Zeitalters: Investmentchancen entlang der technologischen Wertschöpfungskette

September 2023

AUTOREN
Jan Beckers
Marcel Oldenkott

whitepaper@bitcap.com

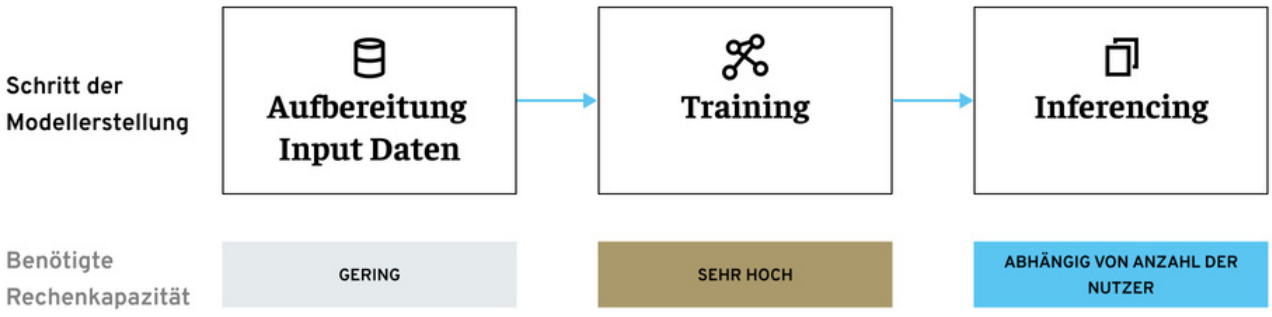
ABSTRACT

Im ersten Teil unserer Whitepaper-Reihe "Das Zeitalter der Künstlichen Intelligenz: Die größte Investmentchance der kommenden Dekade" haben wir gezeigt, wie die Aktienmärkte im ersten Halbjahr 2023 auf Generative Künstliche Intelligenz (GenKI) als zukünftige Technologieplattform reagiert haben. Anhand von Sektor- und Unternehmensbeispielen wurde analysiert, welche Unternehmen von diesem Technologiesprung profitieren dürften und warum. In diesem zweiten Whitepaper liegt der Fokus auf den Infrastrukturunternehmen, die KI-Applikationen ermöglichen.

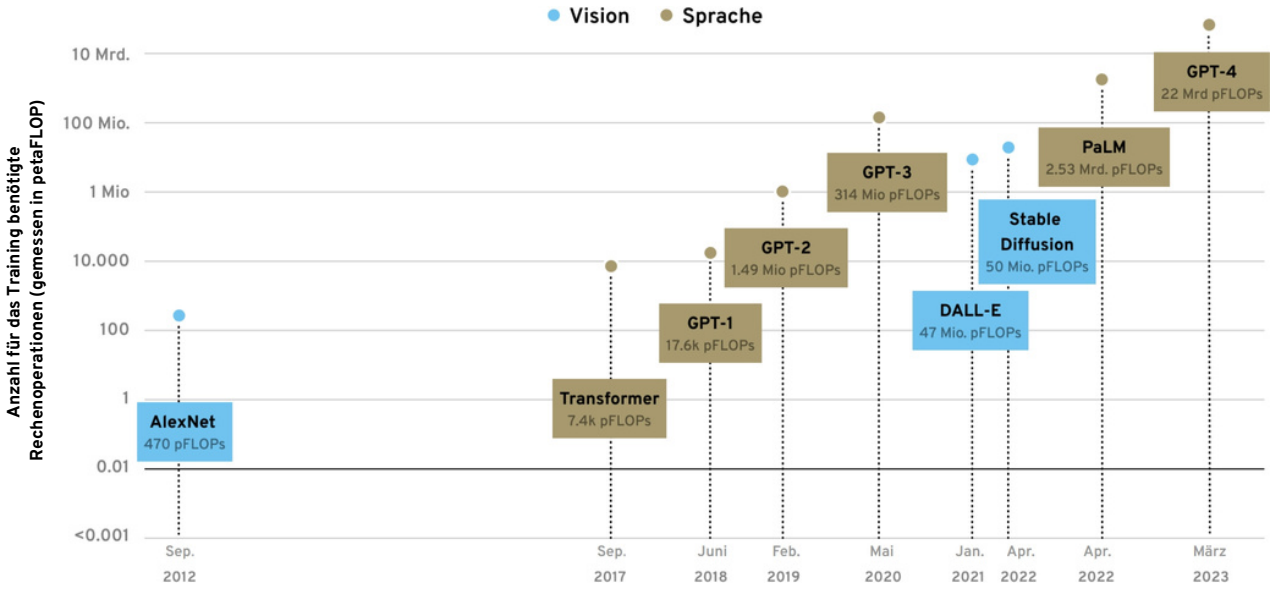
Zunächst wird dargestellt, wie die breite Anwendung von GenKI die Anforderungen an die technologische Infrastruktur erhöht und zu einer steigenden Nachfrage nach KI-Rechenkapazitäten führt. Anschließend wird anhand des Nvidia DGX-Servers die Wertschöpfungskette entlang der Hardware-Komponenten dargestellt und analysiert, welche Unternehmen in diesem Umfeld langfristig zu den Gewinnern zählen könnten.

Künstliche Intelligenz erhöht Infrastrukturanforderungen

GenKI-Modelle basieren auf der Technologie des maschinellen Lernens, die als "mehrschichtiges neuronales Netz" bezeichnet wird und im Wesentlichen versucht, die Funktion des menschlichen Gehirns während des Lernens nachzuahmen. Beim Training werden dem Modell anwendungsspezifische Daten wie Texte oder Bilder zur Verfügung gestellt, die das neuronale Netz analysiert, um daraus komplexe Zusammenhänge zu erlernen. Auf Basis der erlernten Muster kann das Modell anschließend Inferenzen treffen, also Schlussfolgerungen ziehen und Prognosen erstellen. Ein ChatBot kann somit ableiten, welche Aneinanderreihung von Wörtern wahrscheinlich eine zufriedenstellende Antwort auf eine eingegebene Frage ergibt.



Während die Aufbereitung der Input-Daten nicht besonders rechenintensiv ist und mit herkömmlichen Computerchips bewältigt werden kann, verhält es sich bei den Prozessschritten Training und Inferencing anders. Dabei steigen die Anforderungen an die Infrastruktur mit der Komplexität und Größe der Modelle. So benötigen heutige Large Language Models (LLMs) deutlich mehr Trainingsdaten und Rechenkapazitäten als frühere KI-Modelle, wie die folgende Abbildung zeigt.



Quelle: Our World in Data

Betrachtet man beispielsweise das neuronale Netz hinter OpenAI's GPT-4, so wurde dieses mit einem Datensatz trainiert, der 10 Billiarden Wörter umfasst. Insgesamt waren 22 Quadrillionen (10^{24}) Rechenoperationen erforderlich, um alle Zusammenhänge zu erlernen. Der Einsatz herkömmlicher Computerchips ("Central Processing Units" bzw. CPUs), die sequenziell eine Rechenoperation nach der anderen durchführen ("Sequential Computing"), kann einen solchen Anspruch an die Rechenleistung nicht erfüllen. Daher werden die Berechnungen bei GenKI-Anwendungen durch die Methode des "Accelerated Computing" parallelisiert und somit beschleunigt. Dabei dient die CPU hauptsächlich als Steuereinheit, während rechenintensive Spezialaufgaben an dafür optimierte Halbleiter, sogenannte "Accelerators" (Beschleuniger), delegiert werden. Die am weitesten verbreiteten Accelerators sind "Graphical Processing Units", kurz GPUs. Jede GPU verfügt über mehrere tausend Rechenkerne, die datenintensive Berechnungsschritte von GenKI-Modellen simultan durchführen können. Dadurch reduziert sich die benötigte Rechenzeit erheblich.

Obwohl die Rechenkapazität von GPUs die von CPUs deutlich übertrifft, reicht der Einsatz eines oder weniger dieser Chips bei weitem nicht aus. Für das Training und den Betrieb von GenKI-Modellen werden Cluster aus mehreren tausend miteinander verbundenen GPUs benötigt. Für das Training von GPT-4 wurden beispielsweise 25.000 Nvidia-GPUs des Typs A100 für 100 Tage eingesetzt. Dies bedeutet bei Kosten von etwa 1 USD pro GPU pro Stunde implizite Trainingskosten von 60 Millionen USD. Der Betrieb von ChatGPT erfordert bei der aktuellen Anzahl von 132 Mio. Nutzern sogar eine noch höhere Anzahl an GPUs. Schätzungen zufolge kostet der Betrieb täglich 700.000 USD.[1]

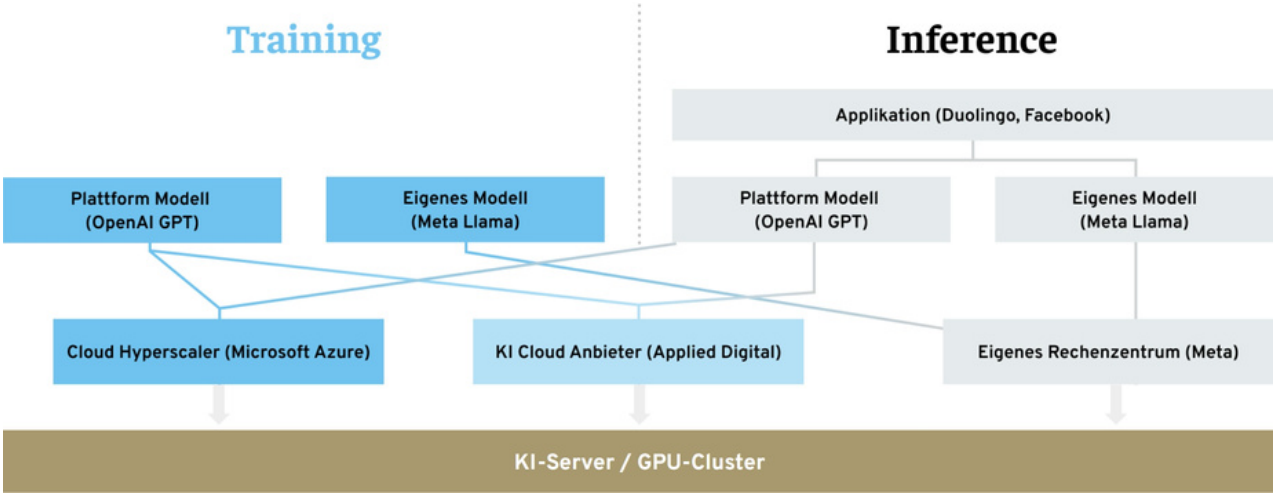
[1] Quelle: SemiAnalysis

Eine Modellplattform wie OpenAI hat drei Möglichkeiten, KI-Rechenkapazität zu nutzen:

a) Betrieb eines eigenen Rechenzentrums
b) Anmietung der Kapazität bei einem Cloud Hyperscaler
c) Anmietung der Kapazität bei einem auf KI spezialisierten Cloud-Anbieter

Die folgende Grafik zeigt, wie bzw. von wem Rechenkapazität für Training und Inferenz nachgefragt wird und welche Unternehmen diese zur Verfügung stellen.

Veranschaulichung der Nachfrage nach KI-Rechenkapazitäten



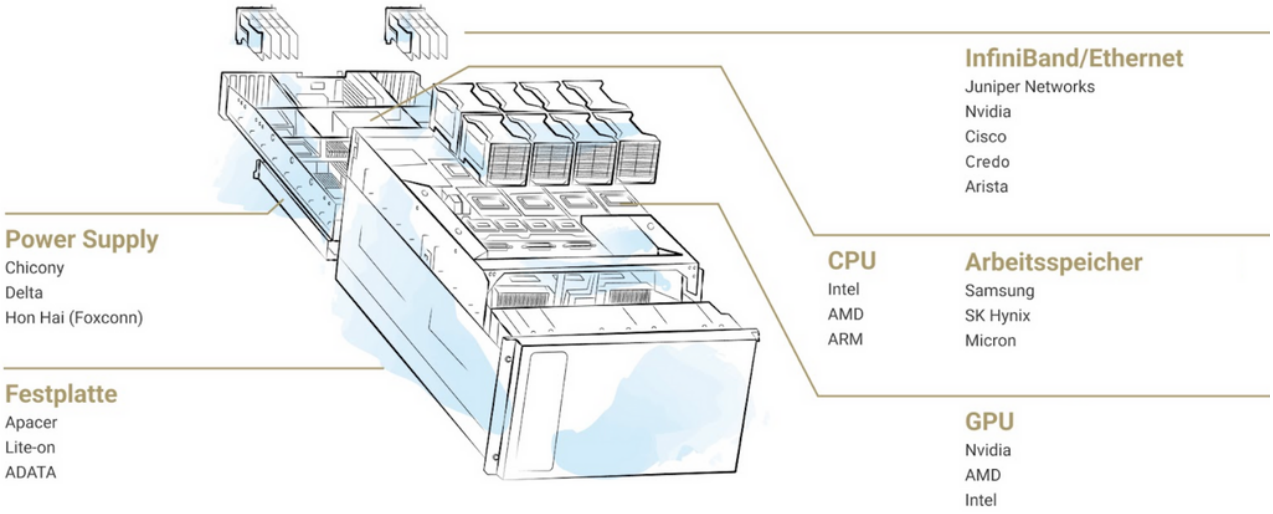
Es wird deutlich, dass die Nachfrage nach KI-Rechenkapazität hauptsächlich auf der Ebene der Modellentwickler und Plattformen entsteht. Wenn beispielsweise ein Endnutzer den KI-Tutor der Sprachlern-App Duolingo nutzt, läuft die Anfrage über OpenAI. Diese lagert die erforderlichen Berechnungen an einen Cloud-Anbieter wie Microsoft Azure aus und verbraucht so KI-Rechenleistung. Unabhängig vom Nutzer der KI-Anwendung oder dem Anbieter der Modelle: Für die Berechnungen wird letztlich immer spezialisierte Hardware benötigt. Im Folgenden erläutern wir am Beispiel des Nvidia DGX Servers die für GenKI zentralen Hardwarekomponenten. Dabei betrachten wir die führenden Hersteller und analysieren, wer zu den langfristigen Gewinnern in diesem Feld gehören könnte.

Investmentopportunitäten im wachsenden KI-Server-Markt

Im vorherigen Abschnitt, haben wir zwei wesentliche Feststellungen getroffen:

a) Die für KI-Prozesse benötigte Rechenkapazität wird von Cloud-Anbietern und unternehmenseigenen Rechenzentren großer Unternehmen bereitgestellt.
b) Die Funktionsfähigkeit von GenKI-Modellen kann nur durch Accelerated Computing gewährleistet werden.

Daraus folgt, dass alle Anbieter auf spezielle KI-Server zurückgreifen müssen, die in dafür vorgesehenen Rechenzentren betrieben werden. Die folgende Abbildung zeigt die schematische Darstellung eines solchen KI-Servers und seiner wesentlichen Hardwarekomponenten sowie die marktführenden Hersteller in diesen Bereichen.



Der KI-Server unterscheidet sich in seiner Hardware-Zusammensetzung und damit auch im Preis erheblich von herkömmlichen Servern. Folgende Unterschiede lassen sich entlang dieser Komponenten feststellen:

Komponente	Traditioneller Server	DGX KI-Server
Anzahl der GPUs	0	8
GPU-Arbeitsspeicher (HBM)	0	8 x 80GB
System-Arbeitsspeicher (DRAM)	1.000 GB	2.000 GB
Netzwerkkarten	1	bis zu 8
Anschlusstechnik	Ethernet	InfiniBand
Anschaffungskosten	**$ 11.000**	**$ 300.000**

Wir gehen davon aus, dass im Zuge wachsender Anwendungsmöglichkeiten von GenKI auch die Nachfrage nach KI-Rechenkapazität rasant ansteigen wird. Die Verschiebung vom Sequential Computing zum Accelerated Computing wird die Marktpositionierung der Unternehmen entlang der Halbleiter-Wertschöpfungskette in den kommenden Jahren maßgeblich beeinflussen. Aktuell werden jährlich rund 14 Mio. Server mit einem Gesamtwert von rund 120 Mrd. USD ausgeliefert. Während für traditionelle Server eine Wachstumsrate von 7% in den nächsten 5 Jahren prognostiziert wird, erwarten Branchenexperten für den Markt mit KI-Servern ein Wachstum von 30-50% p.a. – mit bis zu 150 Mrd. USD Jahresumsatz in 2028.[2]

Das Marktwachstum für KI-Server wird maßgeblich von der Adoption der Technologie auf Anwenderebene abhängen und ist daher schwer einzuschätzen. Momentan fließt ein Großteil der Ressourcen in das Training von KI-Modellen. Zukünftig könnte sich das Inferencing jedoch zu einem größeren Markt entwickeln. Sollten mehrere erfolgreiche KI-Anwendungen wie ChatGPT mit hunderten Millionen aktiven Nutzern entstehen, wird die benötigte Rechenleistung die derzeit verfügbare Kapazität um ein Vielfaches übersteigen. Von diesem Wachstum werden diverse Unternehmen entlang der Wertschöpfungskette profitieren. Im Folgenden werden einige der wichtigsten Anbieter in diesem Ökosystem vorgestellt.

Graphical Processing Units: Der Hochleistungsmotor für GenKI

Der Halbleiterhersteller Nvidia liefert mit seinen GPUs die zentrale Komponente eines KI-Servers und steht daher im Zentrum der steigenden Nachfrage nach KI-Rechenleistung. Nvidias Produktpalette und Integration in das KI-Ökosystem gehen jedoch weit über GPUs hinaus. Das Unternehmen ist daher aus unserer Sicht ideal positioniert, um die führende Plattform für Accelerated Computing und KI-Anwendungen zu werden. Die folgende Grafik zeigt die vertikale Integration von Nvidia entlang der Hauptgeschäftsfelder: Hardware, Software und Anwendungen.

Der Nvidia-Stack im Überblick

Anwendungen	**Plattformen** Nvidia HPC, Nvidia Omniverse, Nvidia AI	**Vortrainierte Modelle** NGC Catalog, Megatron 530B LLM	**Applikationen** NVIDIA NeMo, NVIDIA Picasso, NVIDIA BioNeMo
Software	**Programmiersprache** CUDA	**Libraries** cuDNN, cuBLAS, TensorRT	**Toolkit** Compiler, Debugging, Cluster Management, Monitoring
Hardware	**Chips** GPU, CPU, DPU	**Netzwerk-Technologien** InfiniBand, NVLink, NVSwitch	**Server-Spezifikationen** DGX, HGX

Quelle: Nvidia

[2] Quelle: Kommentar von Lisa Su, CEO AMD, Juni 2023

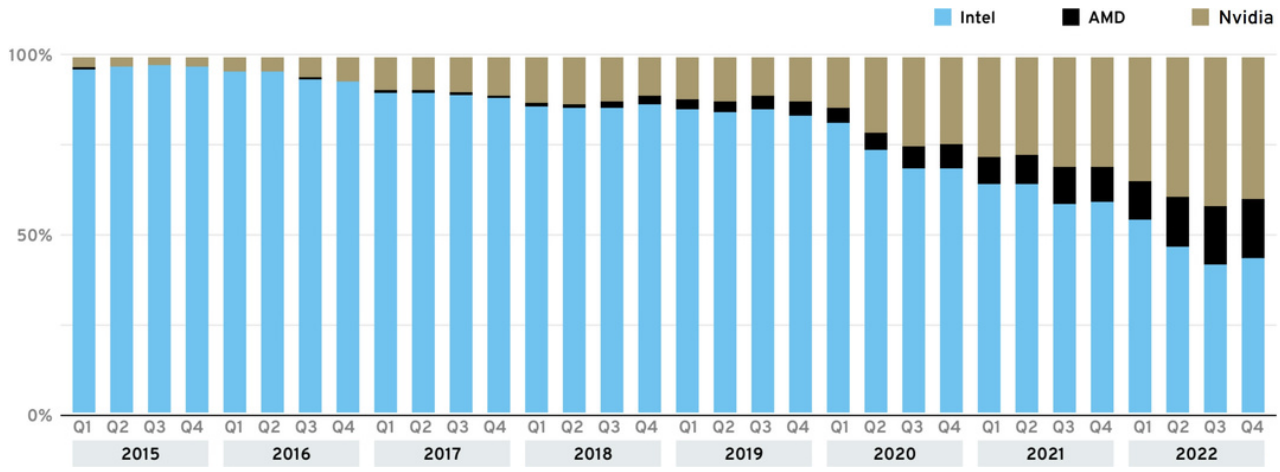
Die umfangreiche **Hardware**-Palette bildet das Fundament des Unternehmens. Die GPUs stellen den Kern des Angebots dar, das auch diverse Netzwerktechnologien sowie voll ausgestattete DGX KI-Server umfasst. Nvidia ist dadurch in der Position, das gesamte "System" zu optimieren, um ein optimales Zusammenspiel aller Hardwarekomponenten zu ermöglichen. Möchte ein Kunde beispielsweise ein LLM trainieren, müssen tausende GPUs miteinander zu einer steuerbaren Einheit bzw. einem Cluster verbunden werden. Nvidias Netzwerklösungen enthalten alle notwendigen Komponenten – vom eigenen Netzwerkchip (DPU) über die Hardwareschnittstelle InfiniBand bis hin zu Technologien wie NVLink und NVSwitch, die eine Kommunikation zwischen mehreren GPUs innerhalb eines Servers oder zwischen mehreren Servern ermöglichen.

Doch nicht nur die Netzwerktechnologie entscheidet über die Leistungsfähigkeit eines KI-Servers oder Clusters, auch die **Software** spielt eine entscheidende Rolle. Mit CUDA bietet Nvidia eine eigene Programmiersprache für GPU-basierte Software, um die sich ein eigenes Ökosystem gebildet hat. Vorgefertigte Funktionen, Entwicklungskits und Schnittstellen beschleunigen nicht nur den Entwicklungsprozess, sondern gewährleisten durch den nahtlosen Übergang zwischen Software und Hardware auch eine höhere Performance. CUDA hat sich seit seiner Einführung im Jahr 2007 zu einem der entscheidenden technologischen Wettbewerbsvorteile des Unternehmens entwickelt. Für Wettbewerber wird es mittelfristig kaum möglich sein, mit der Tiefe und Breite des Softwareangebots und der Entwickler-Community von Nvidia zu konkurrieren.

Basierend auf Nvidias Hardware und Software-Plattformen werden schließlich holistische, **domänenspezifische Lösungen** angeboten. Das Angebot "NVIDIA AI" ermöglicht die direkte Nutzung sowie die Anpassung spezifischer Modelle für verschiedene Anwendungsfälle in Bereichen wie Text (NVIDIA NeMo), Visualisierung (NVIDIA Picasso) oder in der Biologie (NVIDIA BioNeMo). NVIDIA BioNeMo beispielsweise ist ein Cloud-Service für die Anwendung von GenKI in der Arzneimittelforschung. Wissenschaftler können damit GenKI-Modelle für die biomolekulare Forschung in großem Maßstab nutzen, um beispielsweise die Eigenschaften und Funktionen von Molekülen vorherzusagen und so die Entwicklung neuer Arzneimittel zu beschleunigen.

Auf Basis dieser tiefen vertikalen Integration konnte Nvidia bereits vor dem Durchbruch der GenKI zunehmend Marktanteile an den Ausgaben für Cloud-Datenzentren gewinnen (siehe Grafik). Mit der fortschreitenden Adoption von GenKI wird Nvidia seinen Marktanteil weiter ausbauen können.

Nvidia gewinnt durch GenKI Marktanteile an den Ausgaben für Cloud-Datenzentren



Quelle: BIT Capital Research

Betrachtet man vergleichbare, substanzielle Evolutionsstufen der Verbreitung und Entwicklung von Computern, so waren es stets einzelne Plattformanbieter, die die jeweilige Ära durch ihren technologischen Vorsprung und ihre Integration in das Ökosystem dominierten und dabei rund 80% der Profite für sich beanspruchten.[3] IBM gelang dies mit Mainframes, Microsoft und Intel bei der Einführung des PCs, Apple und Qualcomm im Smartphone-Markt. Gegenwärtig sehen wir Nvidia in der besten Position, der dominierende Plattformanbieter in der Ära der Künstlichen Intelligenz zu werden.

High-Bandwidth-Memory: Der kritische Engpass für KI-Arbeitsprozesse

Neben den GPUs ist der Arbeitsspeicher eine weitere Hardwarekomponente, die entscheidend für die Geschwindigkeit eines KI-Servers ist. Zur Durchführung von Trainingsprozessen müssen die benötigten Trainingsdaten und Modellparameter in den Systemarbeitsspeicher des Servers geladen werden. So können die für die geplanten Teilberechnungen benötigten Daten möglichst schnell an die GPUs übertragen werden. Diese Übertragungsgeschwindigkeit wird in GB/Sekunde gemessen und auch als Bandbreite bezeichnet. Bei geringer Bandbreite steigt die Dauer, in der die GPU Daten liest und schreibt, zu Lasten der eigentlichen Berechnungszeit. Somit sinkt der Auslastungsgrad der GPU, wodurch die Kosten für die Berechnung steigen. Der Arbeitsspeicher ist daher ein kritischer Engpass für KI-Arbeitsprozesse.

Eine wesentliche Innovation beim Accelerated Computing besteht darin, dass die GPU nicht jedes Zwischenergebnis in den zentralen Systemarbeitsspeicher schreibt und von dort wieder neu lädt, sondern dafür einen zusätzlichen, dedizierten Arbeitsspeicher besitzt. Dieser GPU-interne Arbeitsspeicher ist zwar deutlich kleiner als der Systemarbeitsspeicher, hat aber eine weitaus höhere Bandbreite und wird daher auch als High-Bandwidth-Memory (HBM) bezeichnet. So hat Nvidia für die leistungsstärksten GPUs die HBM-Technologie kontinuierlich vorangetrieben, um eine möglichst hohe Nutzungseffizienz des Chips zu gewährleisten.

Vor dem Durchbruch der LLMs waren HBM-Arbeitsspeicher jahrelang ein Nischenprodukt, weil sie nur für Spezialanwendungen relevant waren. Dennoch hat sich die Technologie und damit die Bandbreite konstant weiterentwickelt. Die immer komplexeren GenKI-Modelle haben die Anforderungen an HBM weiter erhöht. Doch nicht nur die Leistung, sondern auch der Komponentenpreis dürfte steigen. Branchenexperten gehen davon aus, dass der Markt für HBM an der Schwelle zu einer breiten Marktdurchdringung steht und das Potenzial hat, bis 2027 um das 2 ½-fache zu wachsen. Dies entspräche einem Umsatz von 5,2 Mrd. USD jährlich.[4]

Typ	HBM			
	HBM2	HBM2e	HBM3	HBM3e
Jahr	2016	2018	2022	2024
Bandbreite (GB/s)	256	460	819	1150
Input/Output Geschwindigkeit (Gbps)	2.0	3.6	6.4	8.4
Geschätzter Preis pro GB ($)	10	12	16	20

[3] Quelle: Jefferies [4] Quelle: Gartner

Sowohl der Systemarbeitsspeicher als auch der HBM bestehen aus DRAM (Dynamic Random Access Memory). Im Falle des HBM sind mehrere DRAM-Einheiten übereinander gestapelt und über einen Silizium-Imposer verbunden. Dadurch können anstatt der 64 Datenleitungen eines herkömmlichen Memory-Chips über tausend pro Chip untergebracht werden, wodurch die Speicherbandbreite deutlich erhöht wird. Unserer Ansicht nach sind die globalen Marktführer im Arbeitsspeicher-Bereich, Samsung, SK Hynix und Micron, am besten positioniert, auch im HBM-Markt eine führende Rolle zu übernehmen.

Micron ist nach Samsung und SK Hynix mit einem Marktanteil von rund 25% der drittgrößte Akteur auf dem 136 Mrd. USD umfassenden DRAM-Markt. Das US-amerikanische Unternehmen gilt als Technologieführer für DRAM und kann daher trotz Volumen- und Skalennachteilen gegenüber größeren Anbietern eine höhere Kosteneffizienz operationalisieren. Auf dem HBM-Markt ist Micron ein "Fast Follower" hinter SK Hynix. Micron hat die Technologie bereits vor einigen Jahren entwickelt und forciert nun angesichts des enormen Wachstumspotenzials durch KI-Server den Ausbau. Das Unternehmen hat kürzlich sein HBM3e-Produkt vorgestellt, das bereits im ersten Quartal 2024 in Serie produziert werden könnte. Obwohl Micron derzeit noch keine HBM-Umsätze generiert, könnte dieses Segment in den kommenden Jahren zu einem zusätzlichen Umsatztreiber werden.

Der Anteil von HBM am jährlichen Gesamtumsatz des globalen Markts für Arbeitsspeicher liegt aktuell im niedrigen einstelligen Prozentbereich. Aufgrund des höheren Preises und des schnell wachsenden Volumens kann HBM jedoch in den nächsten Jahren einen Umsatzanteil im hohen einstelligen Prozentbereich erreichen. Mit dem erwarteten Wachstum des KI-Server-Marktes könnte die Nachfrage leicht die aktuell verfügbaren Produktionskapazitäten übersteigen. Wir sehen Micron aufgrund seines technologischen Vorsprungs und seiner Kostenvorteile am besten positioniert, um von diesem möglichen Nachfrageüberhang überproportional zu profitieren.

Networking: Netzwerktechnologien heben Datenübertragungsraten auf neues Niveau

Für den Betrieb von GenKI-Modellen sind Rechencluster mit mehreren tausend GPUs notwendig. Während die Bandbreite des Arbeitsspeichers bestimmt, wie schnell eine GPU auf die im KI-Server bereitgestellten Daten zugreifen und diese umschreiben kann, bestimmen Netzwerktechnologien die Datenübertragungsraten zwischen KI-Servern innerhalb eines Rechenclusters. Auch hier gilt es, die Daten schnellstmöglich durch optimale Vernetzung zur Verfügung zu stellen, um die hochpreisigen GPUs maximal auszulasten.

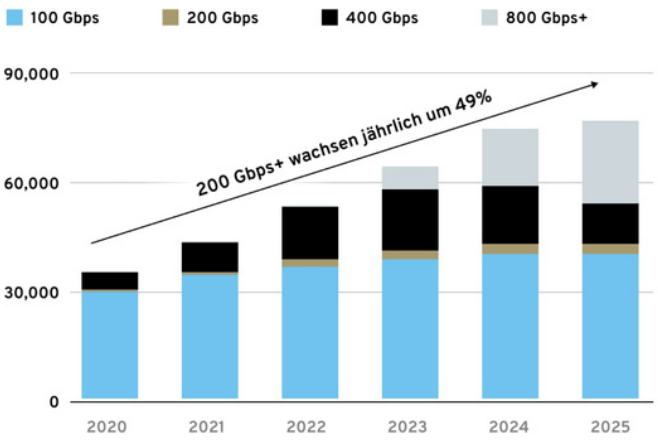
Diese Netzwerktechnologien beinhalten neben physischen Kabeln auch Produkte wie Switches, die einzelne Server zu einem Netzwerk verbinden. Die auf InfiniBand basierenden Netzwerklösungen von Nvidia sind zwar speziell auf KI-Prozesse optimiert und somit in der Datenübertragungsgeschwindigkeit überlegen, jedoch auch fünfmal teurer als Alternativen. Der bevorzugte Branchenstandard ist Ethernet, das im Gegensatz zu Nvidias Infiniband “open source” verfügbar ist und damit mehr externe Innovationen zulässt.

Um auch mit Ethernet-basierten Lösungen höhere Datenübertragungsraten zu ermöglichen, werden in modernen KI-Rechenzentren vermehrt Aktive Elektrische Kabel (AEC) eingesetzt. Diese ersetzen konventionelle Kupferkabel und kombinieren das eigentliche Kabel mit Chip und Software, um das optische Signal bei der Umwandlung in ein elektrisches Signal zu verstärken und zu stabilisieren. Dadurch gehen weniger Daten verloren, sodass Signale nicht mehrfach gesendet werden müssen, bis alle Datenpakete vollständig sind.

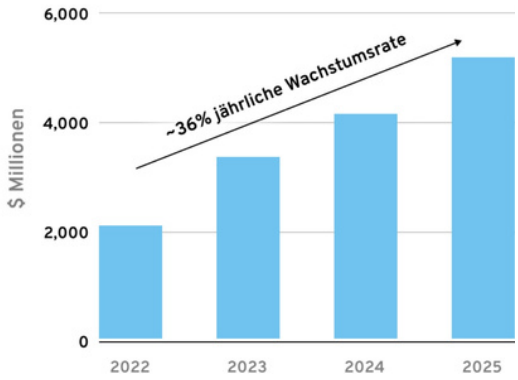
Die Stabilisierung der Datenübertragung lässt höhere Datentransferraten von über 400 Gbit/s zu, erhöht die Energieeffizienz und erlaubt dünnere Kabel. Dies ist gerade bei der Vernetzung von Servern über kurze Strecken von Vorteil. Das 2008 gegründete Halbleiterunternehmen Credo ist der führende Spezialist für AEC-Lösungen.

Während die bisher üblichen Datenübertragungsraten von 100-200 Gbit/s für traditionelle Applikationen ausreichen, benötigen KI-Anwendungen höhere Datentransferraten von bis zu 800 Gbit/s. Um dieser neuen Anforderung gerecht zu werden, werden signifikante Investitionen in die Netzwerktechnik notwendig, was die breite Nachfrage nach den AEC-Produkten von Credo insgesamt deutlich steigern dürfte. Wir gehen davon aus, dass der für Credo relevante Markt für 400-800 Gbit/s-Lösungen mit 36% pro Jahr wachsen und ein Volumen von 5 Mrd. USD im Jahr 2025 erreichen wird.

Anzahl ausgelieferte Ports (in Tsd.)



Prognostiziertes AEC-Marktwachstum

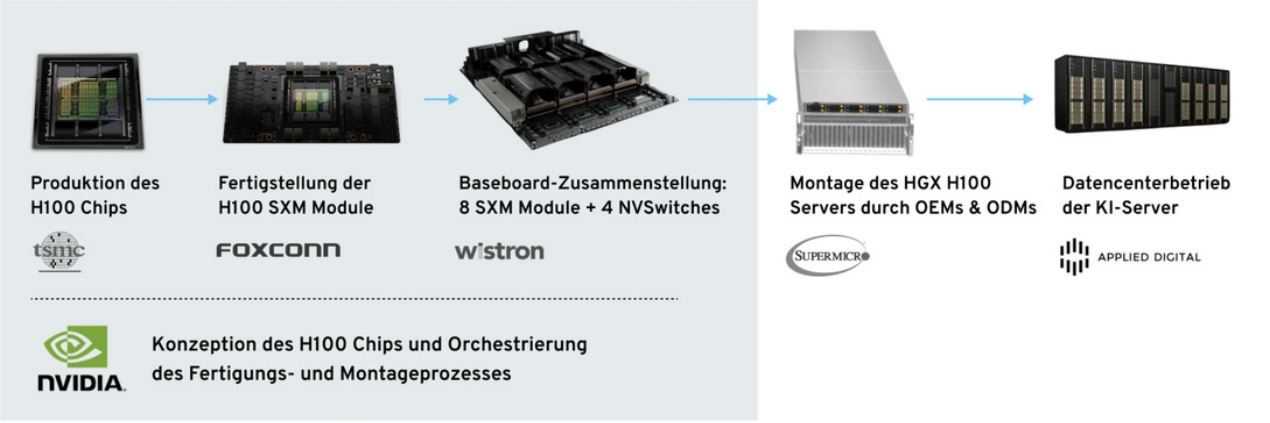


Quelle: Gartner

Neben dem derzeitigen Hauptkunden Microsoft Azure, könnte AWS bald ein zweiter potenzieller Hyperscaler-Kunde werden, der die AEC-Technologie des Unternehmens implementiert. Wir erwarten, dass der KI-Schub das Interesse weiterer großer Abnehmer erheblich verstärken wird und das Wachstum des Unternehmens in den kommenden Jahren positiv beeinflussen könnte. Damit könnte sich Credo von einem Nischenunternehmen zu einem zentralen Anbieter für KI-Kerninfrastruktur entwickeln.

Von Server-Komponenten zum leistungsstarken KI-Rechenzentrum

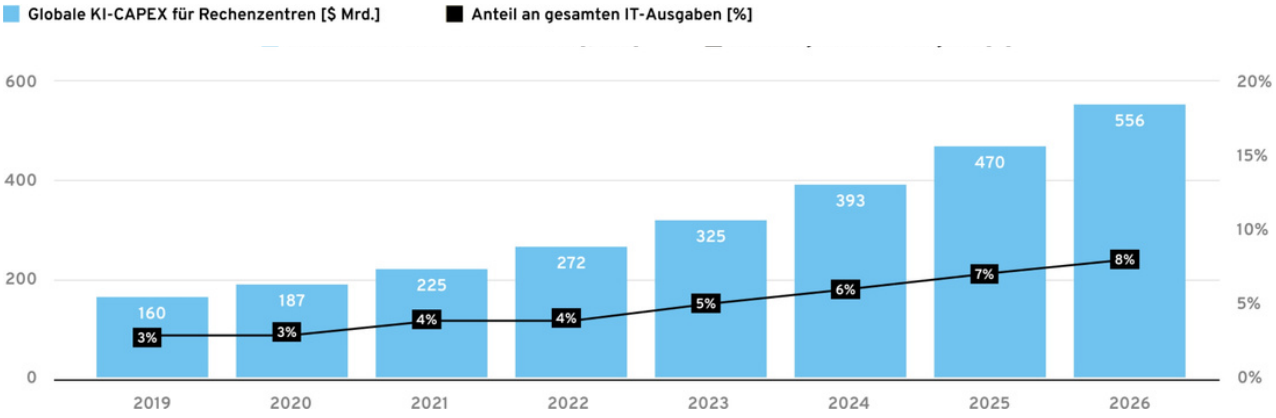
Der aktuelle Wachstumsschub von Nvidia wird vor allem durch die steigende Nachfrage nach KI-Rechenleistung in deren Datacenter-Segment getrieben. Die Betreiber von Datenzentren beziehen einen KI-Server in der Regel jedoch nicht direkt von Nvidia selbst, sondern primär über Serverhersteller wie SuperMicro, die den kompletten KI-Server inklusive GPUs, Arbeitsspeicher, Netzwerklösungen und weiteren Komponenten wie Stromversorgung und Kühlung zusammenstellen.



Die durch GenKI verursachte Verschiebung hin zum Accelerated Computing erhöht die Nachfrage nach spezialisierten Rechenzentren, in denen die Server betrieben werden. GPU-basierte KI-Server verbrauchen mit einer Leistung von 60-80 kW pro Serverschrank ein Vielfaches an Energie pro Stunde im Vergleich zu traditionellen CPU-basierten Servern mit einer Leistung von 10-20 kW pro Serverschrank. Durch die höhere Leistung entwickeln die Server mehr Hitze, wodurch Performance und Haltbarkeit der Hardware beeinträchtigt wird. GPU-basierte Server benötigen somit eine deutlich intensivere Kühlung, die bereits in konventionellen Rechenzentren üblicherweise rund 40% des Stromverbrauchs ausmacht. Spezialisierte Anbieter setzen hier auf innovative Flüssigkühlsysteme, um die Energieeffizienz und die Lebensdauer der Hardware zu verbessern.

Um Kapazitäten für leistungsstarke KI-Server zu schaffen, müssen nicht nur alte Rechenzentren umgerüstet, sondern auch neue errichtet werden. Schätzungen zufolge werden sich KI-bezogene Investitionsausgaben für Rechenzentren innerhalb der kommenden vier Jahre verdoppeln und somit 556 Mrd. USD in 2026 erreichen. Der Zugang zu kostengünstigem Strom und die Erfahrung mit Hochleistungsservern werden bestimmen, wer sich als führender Anbieter in diesem Bereich etabliert.

Verdopplung der KI-bezogenen CAPEX Ausgaben für Rechenzentren bis 2026



Quelle: IDC, Morgan Stanley

Das Unternehmen Applied Digital ist spezialisiert auf den Betrieb von Hochleistungs-Datenzentren. Mit seiner hohen Energieeffizienz und seinen geringen Stromkosten ist es gut positioniert, um geeignete Infrastruktur für KI-Server bereitzustellen. Applied Digital baut aktuell ein Datacenter mit einer Leistung von 200 MW für Accelerated Computing auf und könnte damit bis zu 110.000 Nvidia H100 GPUs betreiben. Zum Vergleich: Mit 21.000 GPUs der älteren A100 Generation verfügte Meta Ende 2022 über einen der weltweit größten GPU-Rechencluster.

Kunden können Nvidia GPUs bei Applied Digital über einen direkten Cloud-Zugang stundenweise oder für mehrere Monate mieten. Das Unternehmen kann mit den im ersten Schritt bestellten 26.000 H100 GPUs bei aktuellen Preisen über 37 Mio. USD Umsatz pro Monat generieren, mit Bruttomargen von mehr als 80%. Der gegenwärtige Kapazitätsengpass bei GPUs ist dabei für derartig hohe Margen verantwortlich.

Als eines von fünf KI-Cloud-Unternehmen in den USA mit Nvidia-Elite-Partner-Status genießt Applied Digital einen besonderen Zugang zu den stark nachgefragten GPUs. Nvidia priorisiert derzeit die Nachfrage von spezialisierten KI-Cloud-Anbietern wie Applied Digital, der die Kapazitäten seiner Rechenzentren für GPUs ausbaut und dabei hochpreisige Komplettlösungen von Nvidia inklusive Netzwerktechnologien erwirbt. Applied Digital trägt damit zu einer breiteren Distribution von Nvidia GPUs und somit zu einer Stärkung des Nvidia-Ökosystems bei. Aufgrund dieser Synergie erwarten wir für die nächsten Quartale ein rasantes Wachstum im KI-Cloud-Segment von Applied Digital.

Zusammenfassung

Die rapide fortschreitende Adoption von GenKI lässt die Nachfrage nach KI-Rechenleistung und damit nach spezialisierten KI-Servern exponentiell ansteigen. Von diesem Nachfrageschub profitieren besonders Infrastrukturanbieter. Die Verschiebung vom Sequential Computing zum Accelerated Computing wird die Marktdynamik in den kommenden Jahren maßgeblich beeinflussen und neue Gewinner hervorbringen.

Nvidia hat sich mit seiner vertikalen Integration als zentraler Plattformanbieter der KI-Ära positioniert und ist damit der dominierende Marktakteur. Aber auch andere Teile der Wertschöpfungskette bieten Chancen für spezialisierte Anbieter. Sowohl der Markt für HBM-Arbeitsspeicher als auch für fortschrittliche Netzwerktechnologien wie AEC werden durch die steigenden Anforderungen von KI-Arbeitsprozessen stark wachsen. Auch bei den Betreibern von Datenzentren werden neue Unternehmen durch Spezialisierung auf Hochleistungsserver Marktanteile für sich erobern können.

Investoren, denen es gelingt, die Gewinner frühzeitig zu identifizieren, können in den kommenden Jahren von überproportionalen Renditen der führenden Unternehmen profitieren. Dabei bleibt das Wachstum nicht nur auf die Ebene einzelner Unternehmen beschränkt. Betrachtet man allein die enorme Kostensenkung bei der Informationsgenerierung durch den GenKI-Schub, lässt sich bereits daraus ein außerordentlicher Produktivitätsgewinn für die gesamte Weltwirtschaft ableiten.

Im dritten Teil unserer Whitepaper-Reihe beleuchten wir die makroökonomischen Implikationen und die konkreten Auswirkungen der Technologie auf die Weltwirtschaft.

Hinweise

Die Verkaufsprospekte und die wesentlichen Anlegerinformationen der von BIT Capital verwalteten Fonds können Sie jederzeit in deutscher und/oder englischer Sprache als PDF auf den jeweiligen fondsspezifischen Unterseiten auf bitcap.com herunterladen. Für Fonds, die keine eigene Unterseite besitzen, finden Sie die Dokumente hier. Die vorgenannten Informationen und Dokumente können auch schriftlich an die BIT Capital GmbH, Dircksenstraße 4, 10179 Berlin, oder per E-Mail angefordert werden. Risikohinweis: Die Fonds weisen aufgrund ihrer Zusammensetzung und des möglichen Einsatzes von Derivaten eine erhöhte Volatilität auf, d. h. die Anteilspreise können auch innerhalb kurzer Zeiträume erheblichen Schwankungen nach oben und nach unten unterworfen sein. Die Zahlenangaben beziehen sich auf die Vergangenheit. Die bisherige Wertentwicklung ist kein Indikator für die zukünftige Wertentwicklung. Aktienkurse können marktbedingt stark schwanken, und somit auch der Fondsanteilswert. Kursverluste sind möglich. Die Fonds investieren in wesentlichem Umfang in Vermögenswerte in anderen Währungen als der Fondswährung. Fällt der Wert dieser Währung gegenüber der Fondswährung, so reduziert sich der Wert des Sondervermögens. Notieren die Fonds in einer fremden Fondswährung, so trägt der Anteilinhaber das Wechselkursrisiko.

Die in dieser Mitteilung enthaltenen Informationen wurden nicht in Übereinstimmung mit den gesetzlichen Bestimmungen über die Erstellung oder Verbreitung von Anlagestrategieberatung oder Anlageempfehlungen erstellt. Diese Kurzinformation ist kein Verkaufsprospekt, sondern dient ausschließlich der Beschreibung ausgewählter Aspekte des Beteiligungskonzepts. Eine Anlageentscheidung kann auf Basis dieser Information nicht begründet werden. Die in dieser Pressemitteilung enthaltenen Informationen stellen weder ein Angebot noch eine Aufforderung zum Kauf oder Verkauf von Anteilen an den von BIT Capital verwalteten Fonds oder anderen erwähnten Finanzinstrumenten dar. Für die Zeichnung ist ausschließlich der gültige Verkaufsprospekt, inklusive etwaiger Nachträge mit den dort fixierten Inhalten, insbesondere der Struktur und den Risiken maßgeblich. Alle Transaktionen mit Anteilen der von BIT Capital verwalteten Fonds basieren auf dem letzten Verkaufsprospekt der Fonds und dem Dokument mit den wichtigsten Informationen für Anleger zusammen mit dem jeweiligen Jahresbericht und/oder Halbjahresbericht der Fonds. Alle Informationen sind sorgfältig zusammengetragen, haben jedoch keinen Anspruch auf Vollständigkeit und sind absolut unverbindlich sowie ohne Gewähr. Des Weiteren dient die Bereitstellung der Information nicht als Rechtsberatung, Steuerberatung oder wertpapierbezogene Beratung und ersetzt diese nicht. Es wird keine Haftung für die Inhalte, welche sich aus dem Verkaufsprospekt bzw. Risiken, die sich aus dem Erwerb des Beteiligungskonzepts ergeben, übernommen. Eine an den persönlichen Verhältnissen des Kunden ausgerichtete Anlageempfehlung, insbesondere in der Form einer individuellen Anlageberatung, der individuellen steuerlichen Situation und unter Einbeziehung allgemeiner sowie objektspezifischer Grundlagen, Chancen und Risiken, erfolgt ausdrücklich nicht. Die hier bereitgestellten Informationen ersetzen keine Anlageberatung, die BIT Capital ansonsten vor jeder Anlage ausdrücklich empfiehlt. Die steuerlichen Auswirkungen für einen Anleger hängen von seinen persönlichen Umständen ab und können sich ändern.

BIT Capital GmbH
Dircksenstraße 4
10179 Berlin

BITCAP.COM