White Paper

Warum die Entwicklung und Bereitstellung von KI-Technologie auf Workstations sinnvoll ist

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

IDC-STANDPUNKT

KI hat sich in allen Branchen als wichtige und differenzierende Fähigkeit rasant etabliert. Ebenso schnell entwickelt sich auch die zur KI-Ausführung benötigte Hardware weiter. Im Fokus der Technologiebranche steht in der Regel das exponentielle Wachstum der meisten modernen KI-Modelle. In den Diskussionen geht es um zig Milliarden Parameter, eine Verringerung der Präzision, die Erweiterung des Speichers, einen High-Performance-Computing (HPC)-ähnlichen Bedarf an KI-Training und -Inferenz sowie Racks mit beschleunigten Servern. In Wahrheit sind diese exorbitant hohen Angaben im KI-Computing die Ausnahme, besonders bei Unternehmen.

Heutzutage arbeiten viele Unternehmen engagiert an der Umsetzung ihrer KI-Initiativen, einschließlich generativer KI, für die keine Supercomputer erforderlich sind. Tatsächlich erfolgt ein Großteil der KI-Entwicklung – und zunehmend auch der KI-Bereitstellung, insbesondere am Edge – auf leistungsstarken Workstations. Sie bieten zahlreiche Vorteile für die KI-Entwicklung und -Bereitstellung. Beispielsweise müssen KI-WissenschaftlerInnen oder -EntwicklerInnen keine Serverzeit mehr aushandeln, denn Workstations stellen die GPU-Beschleunigung bereit. Das ist wichtig, weil serverbasierte GPUs nach wie vor nicht in allen Rechenzentren vorhanden sind. Im Vergleich zu Servern sind sie zudem ausgesprochen kostengünstig und eine Einmalausgabe (anstatt einer schnell ansteigenden Rechnung für eine Cloud-Instanz). Außerdem bieten sie die Gewissheit, dass sensible Daten sicher in On-Premise-Speichern geschützt sind. Somit müssen sich WissenschaftlerInnen und EntwicklerInnen auch keine Sorgen mehr über steigende Kosten machen, wenn sie nur mit KI-Modellen experimentieren möchten.

Bei KI-Bereitstellungsszenarien sieht IDC den Edge in einer dominanteren Funktion als On-Premise- oder Cloud-Instanzen. Auch hier kommt den Workstations als Plattform für die KI-Inferenzierung eine zunehmend wichtigere Rolle zu. Häufig sind nämlich keine GPUs erforderlich, sondern die Inferenz lässt sich auf den softwareoptimierten CPUs ausführen. Die Anwendungsfälle für die KI-Inferenzierung auf Workstations am Edge nehmen schnell zu und umfassen AIOps, Katastrophenhilfe, Radiologie, Erdöl- und Erdgasexploration, Bodenbearbeitung, Telemedizin, Verkehrsregelung, Überwachung des Fertigungswerks und Drohnen.

In diesem Whitepaper wird die zunehmende Rolle von Workstations bei der KI-Entwicklung und -Bereitstellung beleuchtet. Außerdem stellen wir kurz das Dell Portfolio an Workstations für den KI-Bereich vor.

July 2023 IDC-Nr. US51017223

ÜBERSICHT ÜBER DEN STATUS QUO

Explosionsartige Zunahme von KI und Folgen für die Infrastruktur

Die Anzahl der weltweit von Unternehmen vorangetriebenen KI-Projekte nimmt rapide zu. In allen Branchen werden bereits zahlreiche Aufgaben von Software ausgeführt, die in Teilen oder ganz von einem KI-Modell gesteuert ist. IDC verfolgt KI auf vielen Ebenen. Eine der zu berücksichtigenden Kennzahlen ist der prognostizierte Betrag, den Unternehmen und Cloud-Serviceanbieter in Server für die KI-Entwicklung und -Ausführung investieren. Bis 2026 werden das 34,6 Mrd. USD sein, also rund 22 % der weltweiten Ausgaben für Server.

Allerdings besteht das Gesamtbild nicht nur aus Servern. Ein Großteil der Vorbereitung, der Entwicklung, des Prototyping und, in immer höherem Maße, der *Bereitstellung* von KI erfolgt auf Workstations. Sowohl kleine als auch Großunternehmen haben erkannt, dass sie neue Geschäftsmöglichkeiten realisieren können, indem sie ihre Anwendungen mit einigen KI-Funktionen erweitern. Seither sind Experimente mit KI-Modellen sprunghaft angestiegen – und robuste Workstations eignen sich mit ihrer sofortigen Verfügbarkeit und ihrer unmittelbaren Nähe zu den Daten ideal für diesen Zweck.

Wie konnte KI so plötzlich so vorherrschend werden? Immerhin werden KI-Algorithmen bereits seit Jahrzehnten bereitgestellt. Das ist in erster Linie darauf zurückzuführen, dass zwei essenzielle Bedingungen für einen besonders erfolgreichen KI-Algorithmustyp, nämlich das neuronale Netzwerk, in den letzten Jahren realisiert werden konnten: die einfache Verfügbarkeit von schier unerschöpflichen, billigen und unterschiedlichen Datentypen (wie z. B. unstrukturierte und semistrukturierte Daten) und die Erweiterung von linearem Compute durch ein paralleles Modell, um diese neuronalen Netzwerke in einem akzeptablen Zeitraum verarbeiten zu können. Nachdem diese beiden grundlegenden Anforderungen erfüllt waren, haben Data Scientists enorme Fortschritte bei der Entwicklung neuronaler Netzwerke gemacht, die automatisch „lernen“, wie sie immer komplexere Aufgaben ausführen sollen. Das herkömmliche maschinelle Lernen (ML) ist nach wie vor relevant für textbasierte oder numerische Daten, während Deep Learning (DL) bei Video, Audio, Sprache usw. effektiver ist.

Herkömmliche ML-Modelle können in der Regel auf den CPUs einer Workstation entwickelt werden, die im Bestfall über mehrere Dutzend Cores verfügen. Neuronale Netzwerke hingegen erfordern Coprozessoren für die Parallelverarbeitung auf Tausenden von Cores. Das liegt hauptsächlich daran, dass die Extraktion und Klassifizierung von Merkmalen beim maschinellen Lernen ein manueller Prozess ist. Bei Deep Learning ist dies automatisiert, das Modell muss hier durch konstante Wiederholung anhand von großen Datenvolumen trainiert werden. Der gängigste Coprozessor ist derzeit die GPU, aber Start-ups entwickeln neue KI-spezifische Prozessoren, die bald erhältlich sein werden. Diese Art der Beschleunigung, einen separaten Coprozessor für die Parallelverarbeitung einzusetzen, hat die Server- und Workstation-Märkte revolutioniert und eine Entwicklung vorangetrieben, die IDC als massives Parallel-Computing bezeichnet.

Im Jahr 2022 lag der weltweite Markt für beschleunigte Server bei 21,8 Mrd. USD. Diese Summe wird bis 2026 auf 43,4 Mrd. USD ansteigen, rund 57 % davon werden in beschleunigte Server für die KI-Ausführung investiert. Gleichzeitig stieg die Anzahl der separaten GPUs, die für den Einsatz in Workstations verkauft wurden, im Jahr 2022 auf 6,4 Mio. an. IDC schätzt, dass der Markt für Workstations, die für wissenschaftliche Zwecke oder für die Softwareentwicklung eingesetzt werden – beides zunehmend von der KI-Entwicklung vorangetrieben –, bis 2026 auf etwa 2 Mrd. USD wachsen wird.

KI-Entwicklungsphasen

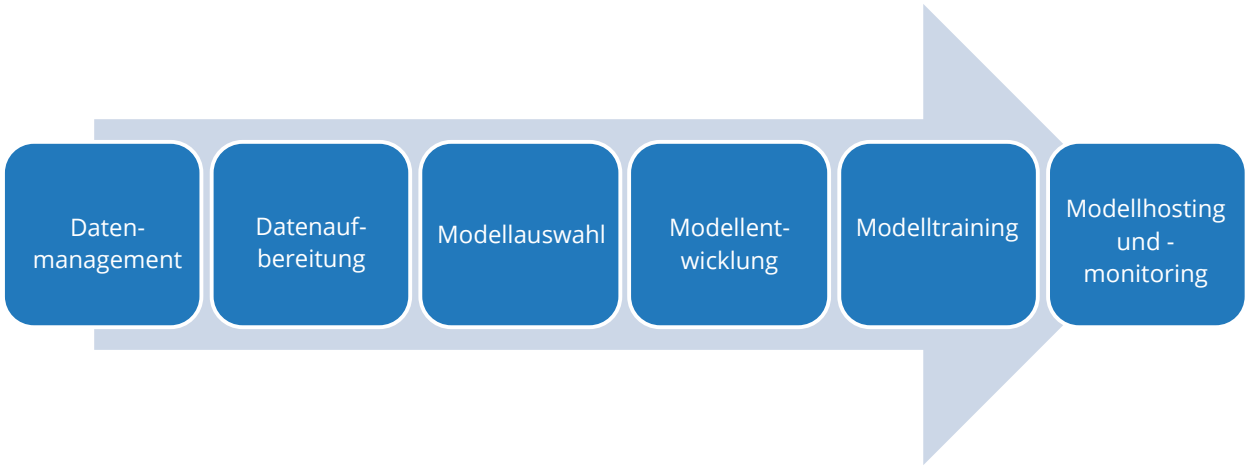
Wie bereits erwähnt ließen sich neuronale Netzwerke aufgrund der zunehmenden Datentypen und -volumes sowie der neuen Compute-Ansätze realisieren. Der erste Teil dieser Gleichung, also Datenvolumes und -typen, ist keineswegs trivial: Manchen Einschätzungen zufolge machen Datenmanagement und aufbereitung bei DL-basierten KI-Initiativen bis zu 80 % des Gesamtaufwands aus. Bevor Modelldesign und -training stattfinden können, müssen die Daten erst zugeführt, verwaltet und aufbereitet werden. IDC unterscheidet folgende KI-Entwicklungsphasen (siehe Abbildung 1):

- **Datenmanagement:** Identifizierung und Verwaltung der für das KI-Modell relevanten Daten aus den schier unerschöpflichen Datenvolumen im Rechenzentrum, am Edge und in der Cloud, die ein Unternehmen sammelt, erzeugt und/oder erwirbt (das kann jeder Datentyp sein, ereignisgesteuert oder gestreamt, und für viele davon ist ggf. eine Art Governance erforderlich)
- **Datenaufbereitung:** Speicherung der (Datei-, Block- oder Objekt-)Daten in einem Data Warehouse oder Data Lake, Bereinigung der Daten, Verifizierung der Vollständigkeit und Qualität der Daten und anschließende Transformation der Daten in ein vom KI-Modell verwertbares Format, z. B. mit Spark oder Tools wie Pandas
- **Modellauswahl:** Entscheidung, welches Modell die KI-Aufgabe, für die es programmiert wurde, am besten ausführen kann (in Bezug auf die Fehlerrate und/oder Performance)
- **Modellentwicklung:** Design des KI-Modells mithilfe von Frameworks wie z. B. XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn oder H2O
- **Modelltraining:** Training des Modells auf der Compute-Infrastruktur mit ausreichend Prozessor- und/oder Coprozessor-Cores für die Parallelverarbeitung – das umfasst zunehmend auch die Fähigkeit, die Entscheidungen eines Modells zu erklären, zu validieren und zu dokumentieren, um Fairness, Verantwortlichkeit und Transparenz sicherzustellen sowie das Prototyping mit Tests des trainierten Modells anhand von Inferenzierung
- **Modellhosting und -monitoring:** Bereitstellung des Modells in einer Produktionsumgebung, damit es die Aufgabe ausführt, für die es entwickelt wurde – dies wird in der Regel als „KI-Inferenzierung“ bezeichnet –, und Monitoring der Modellperformance

Workstations können in Kombination mit dem Rechenzentrum, der Cloud oder der Edge-Infrastruktur in jeder dieser 6 Phasen eine wichtige Rolle übernehmen.

ABBILDUNG 1:

KI-Entwicklungsphasen



Quelle: IDC, 2023.

ENTWICKLUNG VON KI-MODELLEN AUF WORKSTATIONSS

Workstations im Vergleich zu PCs

Es ist allgemein bekannt, dass PCs (Personal Computer) nicht leistungsstark genug für die KI-Entwicklung sind. Data Scientists und KI-EntwicklerInnen arbeiten meist an strategisch wichtigen Projekten für das Unternehmen, sodass die ununterbrochene Produktivität höchste Priorität hat. Workstations sind tendenziell zuverlässiger als PCs. Sie werden in der Regel mit leistungsstärkeren Komponenten gefertigt und sind für die Software optimiert, die auf ihnen ausgeführt wird.

Zu diesen Komponenten gehören:

- **Hochwertige Prozessoren:** Ein Beispiel sind skalierbare Intel Xeon Prozessoren.
- **Leistungsstarke GPUs:** Ein Beispiel sind professionelle NVIDIA RTX-GPUs wie die NVIDIA RTX 6000 Ada.
- **Mehr Storage:** Einige Workstations bieten bis zu 60 TB Storage und tendenziell deutlich höhere I/O-Geschwindigkeiten als PCs.
- **Mehr Arbeitsspeicher:** Workstations sind mittlerweile mit 6 TB Arbeitsspeicher erhältlich.
- **Kühlung:** Leistungsstarke Komponenten erzeugen viel Wärme – Data Scientists benötigen Workstations mit guter Kühlung, um eine Überhitzung zu vermeiden und die optimale Performance sicherzustellen.
- **NIC (Network Interface Card):** Hochgeschwindigkeits-NICs sind entscheidend für die schnelle und effiziente Datenübertragung – besonders für Data Scientists, die mit auf Remoteservern hinterlegten, umfangreichen Datenvolumen arbeiten.
- **Display:** Qualitativ hochwertige Displays sind essenziell für die Datenvisualisierung – Data Scientists benötigen Monitore mit hoher Auflösung, Farbgenauigkeit und großem Bildschirm.

- **ECC-Arbeitsspeicher (Error-Correcting Code):** Die ECC-Funktion erkennt und korrigiert die gängigsten Arten der internen Datenbeschädigung, verhindert Bluescreens bei langen KI-Trainings aufgrund von einem „harten“ (defektes Bit) oder einem „weichen“ Fehler (Bit-Flip, verursacht fehlerhafte Werte) und stellt zudem die Ergebnisgenauigkeit sicher (wichtige Anforderung für lebensentscheidende Tätigkeiten, wie z. B. im Gesundheitswesen).
- **Spezielle Chips:** Ein Beispiel sind Intel Movidius Vision Processing Units (VPUs) – Coprozessoren für paralleles Computing für Computer Vision und KI-Anwendungen am Edge für den Einzelhandel und Bereiche wie Sicherheit und industrielle Automatisierung – oder die Verwendung von FPGAs in Workstations, wie z. B. Finanzanwendungen.
- **Optimierungssoftware:** Beispiele sind OneAPI (standardbasiertes Programmiermodell von Intel für die einfachere Entwicklung und Bereitstellung von datenzentrierten Workloads auf CPUs, GPUs, FPGAs und anderen Accelerators) und CUDA (NVIDIAs Plattform für paralleles Computing und Application Programming Interface für die Ausführung von allgemeinen Workloads auf GPUs).

CPUs im Vergleich zu GPUs für KI

Workstations lassen sich in verschiedenen Phasen der KI-Entwicklung einsetzen und eignen sich meist für zahlreiche Einsatzmöglichkeiten. Trotz der Betonung von GPUs für die Parallelverarbeitung kommt auch den CPUs eine entscheidende Rolle zu, wenn es um die Entwicklung eines KI-Modells auf einer Workstation geht. Ebenso wie GPUs können CPUs für die Datenmanipulation und natürlich für die Entwicklung von herkömmlichen ML-Modellen verwendet werden. CPUs werden zudem für die Datenexploration eingesetzt, den Prozess, bei dem visuelle Darstellungen des Datenvolumens herangezogen werden, um die Merkmale der Daten zu verstehen.

Beim DL-Training sind die Host-CPUs nicht mehr so wichtig, weil die GPUs den eigentlichen Trainingsprozess ausführen. Dennoch fungieren die CPUs hier weiterhin als Verarbeitungsschicht für die kritische Software (wie z. B. das BS oder CUDA) und für die Orchestrierung der Prozesse zwischen den GPUs oder mit anderen Chips. Des Weiteren übernehmen CPUs mehr und mehr die neue Funktion einer KI-Inferenzierungs-Engine, insbesondere wenn für die Ausführung eines KI-Modells in der Produktion eine Workstation verwendet wird. Laut IDC werden im Jahr 2024 die Infrastrukturausgaben für KI-Inferenzierung höher sein als die KI-Infrastrukturausgaben für KI-Training – und ein signifikanter Teil (39 %) dieser Inferenzierung wird auf den Host-CPUs stattfinden.

Workstations im Vergleich zu Servern: eine symbiotische Beziehung

In den meisten Unternehmen ist Pragmatismus angesagt, wenn es darum geht, eine Workstation, einen On-Premise-Server, eine Cloud-Instanz oder eine beliebige Kombination dieser drei für die KI-Entwicklung bereitzustellen. Zwischen den Workstations, Servern und Cloud-Instanzen der verschiedenen Entwicklungsphasen eines KI-Projekts besteht eine symbiotische Beziehung.

Im Vergleich zu Rechenzentrumsservern bieten Workstations den Vorteil, dass Data Scientists überall arbeiten können. Das war nicht nur während der Pandemie ein kritischer Faktor, sondern ist auch unter „normalen“ Umständen von Bedeutung. Zudem können sie frei mit ihren KI-Modellen experimentieren und sie so häufig iterieren, wie sie es für erforderlich halten. Schließlich ist dank der Leistungsfähigkeit moderner Workstations mit leistungsstarken GPUs der iterative Prozess oft interaktiver und liefert sofort Feedback und Resultate. Und zwar, ohne dass Zugriff auf Server angefordert werden muss oder andere Einschränkungen des Rechenzentrums zum Tragen kommen. Außerdem bieten Workstations die Flexibilität, den Computer näher an die Daten zu bringen anstatt umgekehrt. Das spart Bandbreite, verringert die Netzwerküberlastung und erhöht den Durchsatz. Des Weiteren lassen sich Workstations für unterschiedliche Anforderungen konfigurieren, wie z. B. herkömmliche ML-Aufgaben oder eher DL-intensive Tätigkeiten.

Auch wenn es ein erhebliches Wachstum im Marktsegment der beschleunigten Server gibt, sind diese noch nicht in allen Unternehmensrechenzentren verfügbar. Zum Zeitpunkt der Erstellung dieses Whitepapers waren durchschnittlich 4 % der Server in Unternehmensrechenzentren beschleunigt. Das heißt im Umkehrschluss, dass viele Unternehmen nicht über die Mittel verfügen, KI auf bereits verfügbaren On-Premise-GPUs zu entwickeln oder auszuführen. Auch aus diesem Grund sind beschleunigte Workstations eine gute Alternative für die KI-Entwicklung.

Hochbeschleunigte Workstations sind nun für das DL-Training leistungsfähig genug – sofern das KI-Modell nicht übermäßig groß ist –, sodass für das Training keine Server mehr benötigt werden. Modelle, die auf Workstations mit GPUs trainiert wurden, lassen sich entweder auf Workstations oder auf Servern ohne GPUs bereitstellen. In diesem Fall werden die Inferenzfunktionen der CPUs genutzt. Softwaretechnologien wie DL Boost und oneAPI von Intel ermöglichen die KI-Inferenzierung auf der CPU. Damit können auch bereits im Rechenzentrum vorhandene, nicht beschleunigte Server die KI-Anwendungen unterstützen.

Workstations im Vergleich zur Cloud

Cloud-Computing hat die Denkweise von Unternehmen hinsichtlich Infrastruktur, Daten und Anwendungen revolutioniert. Mit dem Versprechen von nahezu unbegrenzter Skalierbarkeit ermöglicht die Cloud den EntwicklerInnen eine On-Demand-Bereitstellung von Ressourcen und beschleunigt potenziell die Innovationsgeschwindigkeit durch weniger Einschränkungen. Scheinbar ist die Cloud das perfekte Paradigma für die KI-Entwicklung.

Das ist jedoch nicht immer so. Eine IDC-Studie hat gezeigt, dass Unternehmen bestimmte Workloads zunehmend wieder aus der Public Cloud in die On-Premise-Infrastruktur zurückholen. Dafür gibt es mehrere Gründe:

- **Cloud-Verfügbarkeit:** Alle, die Cloud-Services nutzen, haben schon einen Ausfall erlebt – entweder aufgrund von Problemen des Cloud-Anbieters selbst oder wegen einer Unterbrechung der Netzwerkverbindung irgendwo zwischen dem Hyperscale-Rechenzentrum und den EndnutzerInnen. In solchen Situationen sind die NutzerInnen vom Serviceanbieter abhängig: Er muss das Problem lösen, während die Produktivität zum Stillstand kommt.
- **Sicherheit und Compliance:** In vielen Branchen diktieren die Policies der Unternehmensführung, wo Daten kommuniziert und gespeichert werden dürfen. Das schränkt die Nutzung von Cloud-Services erheblich ein. Gesetzliche Bestimmungen wie z. B. die DSGVO in Europa und der California Consumer Privacy Act erzwingen ebenfalls Regeln für die Datenhoheit.
- **Kosten:** In der Regel unterschätzen Unternehmen, wie schnell die Gebühren für Cloud-Services ansteigen können. Das gilt insbesondere für Workloads, die Compute-Funktionen mit hoher Performance und viel Storage erfordern. Die Wirtschaftlichkeit der Cloud basiert darauf, dass alle Arten des Ressourcenverbrauchs gemessen werden, darunter auch die Rückführung von Daten in die On-Premise-Infrastruktur.
- **Druck der Trial-and-Error-Methode:** Die meisten KI-Initiativen beginnen mit einer erheblichen Anzahl an Experimenten. Fehlerhafte Modelle sind dabei ein wesentlicher Bestandteil des Entwicklungsprozesses. In diesem Fall zahlen die KI-WissenschaftlerInnen und -EntwicklerInnen einen „psychologischen Preis“, da die Cloud-Rechnung in die Höhe schnellt, sie jedoch noch keine verwertbaren Ergebnisse vorweisen können.

Workstations überwinden diese Einschränkungen und nutzen gleichzeitig cloudnative Technologien, wie z. B. auf Microservices basierende Architekturen und API-gesteuerte Automatisierung. Das bietet einige derselben Vorteile wie beim Vergleich von Workstations mit Rechenzentrumsservern:

- **Überall arbeiten:** Wenn die Abhängigkeit von der Public Cloud entfällt, sind auch isolierte, also nicht verbundene Szenarien wieder möglich. Viele Hochsicherheitsumgebungen sind per Air Gap von öffentlichen Netzwerken getrennt – und KI-Workstations erfüllen genau diese Anforderung. Lokale Ressourcen senken zudem den Bedarf an kostspieligen Netzwerkverbindungen.
- **Datenlokalität:** Die starke Zunahme an IoT-Geräten und anderem vernetzten Equipment trägt zu einem exponentiellen Datenwachstum an Edge-Standorten bei. In vielen Situationen ist es sinnvoll, die Computing-Ressourcen mit einer dedizierten Workstation an einen Colocation-Standort auszulagern. Aufgrund der so eingeschränkten Datenverschiebung lassen sich auch viele Complianceanforderungen erfüllen.
- **Freie Experimente:** Das Training und die Optimierung von KI-Modellen sind iterative Prozesse, die oftmals einige Elemente der Trial-and-Error-Methode beinhalten. EntwicklerInnen brauchen die Freiheit, Experimente ohne den Druck von potenziell höheren Servicegebühren ausführen zu können. Außerdem bieten Workstations mehr Flexibilität für nutzerdefinierte Tools.

Hinsichtlich des letzten Punkts ist der Preisvergleich einer Workstation mit einer Cloud-Bereitstellung relativ einfach, da die meisten Cloud-Serviceanbieter direkt Kostenschätzungen für jede Konfiguration erstellen, die EndnutzerInnen bereitstellen möchten. Beispielsweise belaufen sich die Kosten für eine einzige reguläre VM (virtuelle Maschine) mit einer NVIDIA T4 und einem SSD-Storage von 375 GiB, die 8 Stunden pro Tag an 5 Tagen die Woche genutzt wird, bei einem großen Cloud-Anbieter auf 140 USD. Bei einer Verdopplung der VMs, T4s und SSDs steigen die Kosten auf 365 USD pro Monat. Bei 2 VMs, aber einer Verdopplung auf 4 T4s und 4 x 375 GiB Storage mit einem vollständigen Trainingslauf in der Umgebung schnellen die Kosten in die Höhe: auf 2.700 USD pro Monat. Wir können also mit Fug und Recht sagen, dass die Cloud-Kosten für KI-Entwicklung leicht auf Zehntausende Dollar pro Jahr ansteigen können. Das ist deutlich mehr als der jährliche Wertverlust bei einer High-End-Workstation.

KI-PROTOTYPING AUF WORKSTATIONS

Im Vergleich zu sowohl On-Premise-Servern als auch der Cloud bieten Workstations einen deutlichen Vorteil beim Prototyping von KI-Modellen. Server im Rechenzentrum laufen möglicherweise mit voller Auslastung oder sind zu geschäftskritisch, um für KI-Prototyping und -Tests eingesetzt zu werden. Und wie bereits erwähnt können Cloud-Instanzen schnell das Budget sprengen, wenn sie als Testumgebung fungieren. Mit Workstations müssen KI-WissenschaftlerInnen oder -EntwicklerInnen keine Serverzeit mehr aushandeln oder sich Sorgen machen, dass die Cloud-Rechnungen während der Prototyping-Phase in die Höhe schnellen. Durch ihre niedrigen und einmaligen Kosten bieten Workstations die vollkommene Freiheit, das Prototyping jederzeit und überall ohne zusätzliche Kosten ausführen zu können.

BEREITSTELLUNG VON KI-MODELLEN AUF WORKSTATIONS

Die Entwicklung von KI-Modellen auf Workstations ist seit Jahren eine gängige Strategie. Laut IDC nehmen nun auch die Anwendungsfälle für die *Bereitstellung* von KI-Modellen auf Workstations zu, und das zumeist am Edge. Mit anderen Worten: Das KI-Modell wird in der Produktion auf der Workstation eingesetzt, da dort die Inferenzierung ausgeführt wird. Der Edge wird immer häufiger als KI-Bereitstellungsort für Server genutzt – die jährlichen Hardwareausgaben dafür haben sich von 2020 bis 2024 mehr als verdreifacht. Workstations liegen nicht weit dahinter, da die EndnutzerInnen ihre Vorteile am Edge erkennen.

IDC definiert den Edge als verteiltes Computing-Paradigma, das die Bereitstellung von Infrastruktur und Anwendungen außerhalb der zentralen Cloud und On-Premise-Rechenzentren so nah wie nötig am Standort der Datengenerierung und -nutzung umfasst. Das schließt Remotestandorte und Zweigstellen sowie branchenspezifische Standorte wie z. B. Werke, Lager, Krankenhäuser und Einzelhandelsgeschäfte ein.

Daten- und Compute-intensive Workloads werden zunehmend an On-Premise- oder Edge-Standorten bereitgestellt. Damit sollen Einschränkungen von Public Clouds umgangen werden, beispielsweise die erforderliche Zeit für den Upload umfangreicher Datenvolumen und die variablen Kosten für die Ausführung des KI-Trainings, insbesondere in Situationen, in denen eine erhebliche Anzahl an Data-Science-Experimenten erforderlich ist.

IDC-Untersuchungen zufolge ist der Edge ein schnell wachsendes KI-Bereitstellungsszenario: Im Jahr 2023 investierten Unternehmen 2,9 Mrd. USD in KI-Compute-Funktionen am Edge – diese Summe wird bis 2026 auf 6,9 Mrd. USD steigen (siehe *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC-Nr. US49671722, September 2022). Des Weiteren wird der Edge auch als Option für die Bereitstellung von HPC-Workloads wie z. B. im Engineering- und technischen Bereich immer interessanter: Derzeit investieren Unternehmen nahezu 1 Mrd. USD in diese Workloads am Edge – bis 2027 wird dieser Betrag auf 2,4 Mrd. USD anwachsen (siehe *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC-Nr. US50525123, April 2023). In diesen Bereichen ist die Bereitstellung einer KI-Workstation sinnvoll.

Bei der Bereitstellung eines KI-Modells auf einer Workstation am Edge sind nicht unbedingt High-End-GPUs wie bei der KI-Entwicklung erforderlich. „Leichtere“ GPUs sind in der Lage, die KI-Inferenzierung zu übernehmen, und in etlichen Fällen werden überhaupt keine GPUs benötigt. Dann können CPUs die Inferenzierungsaufgabe angemessen ausführen, insbesondere im Zusammenspiel mit Optimierungen wie Intel DL Boost, einer Reihe an ISA-Funktionen (Instruction Set Architecture) auf Intel Mikroprozessoren für die Beschleunigung von KI-Workloads, einschließlich KI-Inferenzierung. Laut Intel ist mit Intel DL Boost ein 1,45-mal höherer Durchsatz bei der INT8-Echtzeitinferenz mit skalierbaren Intel Xeon Prozessoren der 4. Generation (die Intel DL Boost unterstützen) möglich als bei der vorherigen Generation (BERT-Large SQuAD). Auch aus diesem Grund sind Workstations für die Bereitstellung am Edge geeigneter, denn für Leistung, Mobilität und Kühlung wird weniger Energie benötigt. Dank des niedrigen Stromverbrauchs von 12 W passt Intel Movidius Myriad (M2) gut in dieses Energiekonzept.

Anwendungsfälle für die Bereitstellung von KI auf Workstations

In etlichen Situationen ist die KI-Bereitstellung auf lokal bereitgestellten Workstations eine logische Konsequenz. Gängige Merkmale dafür sind große Volumes an maschinengenerierten Zeitreihendaten sowie unstrukturierte Daten wie Videostreams und Bilder. In einigen Fällen müssen Experten mit Fachkompetenz die KI-Modelle anhand von menschlichen Interpretationen ergänzen.

Beispiele:

- **AIOps:** Da die IT-Systeme an Umfang und Komplexität zunehmen, wird es immer notwendiger, von einem reaktiven Incident-Management zum proaktiven Monitoring zu wechseln. Dies gilt insbesondere, wenn Infrastruktur und Anwendungen an Edge-Standorten verteilt werden, an denen nur wenig oder gar kein technisches Personal arbeitet. Durch die Modellierung einer Baseline der „normalen“ Performance lassen sich Anomalien erkennen und Korrekturschritte automatisieren.

- **Katastrophenhilfe:** Im Fall einer Katastrophe müssen die ErsthelferInnen die Situation schnell bewerten, erforderliches Equipment anfordern und Ressourcen dort bereitstellen, wo die Hilfe am dringendsten benötigt wird. Häufig gibt es in einer solchen Umgebung keine Netzwerkverbindung, sodass eine lokale Workstation erforderlich ist, die Datenfeeds aggregieren, mithilfe von KI-Modellen Rückschlüsse ziehen und die Kommunikation mit dem wichtigsten Personal automatisieren kann.
- **Radiologie:** Fortschritte bei der bildgebenden Technologie haben dazu geführt, dass die Datenmenge aus einem einzelnen Scan erheblich zugenommen hat. Diese muss für eine zeitnahe Analyse vor Ort verbleiben. KI-Modelle, die anhand von Millionen vorhandener Beispiele trainiert wurden, erkennen Muster viel präziser als das menschliche Auge und erhöhen so die Genauigkeit.
- **Erdöl- und Erdgasexploration:** Moderne Unternehmen im Erdöl- und Erdgassektor verwenden eine Kombination aus Telemetrie-, seismischen und Bilddaten, um Reservoire natürlicher Ressourcen aufzuspüren, Bohrstellen auszuwählen und die Leistung des Equipments im Produktionsprozess zu optimieren. Häufig müssen dafür Informationen in Gegenden analysiert werden, in denen nur die teure Kommunikation über Satellit möglich ist.
- **Krebsforschung und Arzneimittelentwicklung:** ForscherInnen in Krankenhäusern und an Universitäten setzen KI und die Verarbeitung von natürlicher Sprache ein, um OnkologInnen dabei zu unterstützten, die effektivste, individuelle Krebsbehandlung für ihre PatientInnen zu finden. Sie kombinieren zudem maschinelles Lernen mit Computer Vision, damit die RadiologInnen die Tumorentwicklung bei den PatientInnen besser nachverfolgen können. Außerdem nutzen die ForscherInnen Algorithmen, um besser zu verstehen, wie sich der Krebs entwickelt und welche Behandlung ihn am effektivsten bekämpft.
- **Bewertung von Versicherungsansprüchen:** Die manuelle Antragsbearbeitung ist arbeitsintensiv und anfällig für menschliche Fehler. Wenn die Anspruchsberechtigung von einer KI bewertet wird, senkt das die Kosten, da sich die SachbearbeiterInnen der Versicherung auf die Fälle konzentrieren können, die eine genauere Untersuchung erfordern. Das erhöht den gesamten Vorgangsdurchsatz, ohne die Genauigkeit zu beeinträchtigen.
- **Telemedizin:** KI beschleunigt die Genesung von PatientInnen durch maßgeschneiderte, individuelle Behandlungspläne, die auf den Echtzeitvitalwerten von Wearables basieren. Die Daten werden mit den bisherigen Patientendaten und einer Wissensdatenbank mit ähnlichen Fällen verglichen. Das ist besonders in ländlichen Gegenden mit einer höheren Abhängigkeit von der Telemedizin wichtig.
- **Sicherheit im Einzelhandel (Diebstahlschutz):** Mithilfe von Echtzeitanalysen von Videostreams wird menschliches Verhalten vorhergesehen, das möglicherweise zu einer kriminellen Handlung führt. Um die Bewegungen von Einzelpersonen in einem Geschäft nachzuverfolgen, müssen in der Regel mehrere Videofeeds berücksichtigt werden. Aufgrund der zeitkritischen Natur bei der Identifizierung eines möglichen Diebstahls sollte dieser Prozess am besten vor Ort ausgeführt werden.
- **Verkehrsregelung:** Die für das Verkehrswesen zuständigen Behörden setzen zunehmend KI für die Koordination der Ampelanlagen und digitalen Verkehrsleitsysteme ein, um den Verkehrsfluss zu verbessern und die BürgerInnen zu schützen. Zur Optimierung der Verkehrsmuster ist eine Kombination der Eingabedaten, darunter Videokameras und Telemetrie aus Straßensensoren, erforderlich.
- **Überwachung des Fertigungswerks:** Für die Werksleitung hat die sichergestellte Aufrechterhaltung kritischer Prozesse sowie die Einhaltung der Produktionszeitpläne oberste Priorität. Folglich sind die vorausschauende Wartung von wichtigem Equipment, eine automatisierte Defekterkennung sowie die Optimierung der Lieferkette sowohl am

Standort als auch außerhalb erforderlich. In diesem Bereich kann KI die menschlichen BedienerInnen unterstützen, um die Produktivität zu steigern und die Einhaltung der Sicherheitsstandards sicherzustellen.

- **Drohnen:** Mithilfe der automatisierten Analyse der von Drohnen aufgenommenen Bilder kann ein ganzes Spektrum an Bedingungen in einem Maß überwacht werden, das zuvor nicht möglich war. Das hat signifikante Auswirkungen auf die Inspektion der Erdgas- und Stromnetzinfrastruktur, auf Versicherungsgutachten, Such- und Rettungsaktionen, auf die Präsizionslandwirtschaft sowie die Hege und Pflege in Fischerei- und Tierschutzgebieten.
- **Alltägliche Büroumgebungen:** Ganz alltägliche Büroumgebungen werden zunehmend mit KI-basierten Produktivitätstools wie Microsoft Copilot verbessert.
- **Erneuerbare Energien:** Standorte von erneuerbaren Energien, wie z. B. Windenergieparks, hydroelektrische Staudämme und Solarfarmen, benötigen Echtzeitüberwachung, Wartung und Datenerfassung mit lokaler Erzeugung und Analyse.

DELL WORKSTATIONS FÜR KI

Dell bietet ein breites Spektrum an Workstations für verschiedene Stufen der KI-Entwicklung und/oder -Implementierung, die alle unter der Marke „Data Science Workstation“ (DSW) erhältlich sind. In diesem Abschnitt werden kurz die technischen Daten vorgestellt, danach gehen wir auf die verschiedenen KI-Profile/-Anwendungen (wie z. B. Data Scientists) und die Vorteile der Dell DSW-Technologie ein. Diese KI-fähigen DSWs wurden speziell für Data Scientists entwickelt. Die neuesten Precision-Data-Science-Workstations nutzen KI-Funktionen für das Finetuning der Geräte und um die Performance der von den Data Scientists am häufigsten verwendeten Anwendungen zu optimieren. Damit können sie ihre wichtigsten Tätigkeiten schneller erledigen. Des Weiteren werden Dell Precision-Workstations von unabhängigen Softwareanbietern getestet und zertifiziert. So ist sichergestellt, dass sie die leistungsstarken Anwendungen unterstützen, die Dell Kunden für ihre täglichen Aufgaben benötigen.

Alleinstellungsmerkmale von Dell Workstations

Dell Precision-Workstations mit NVIDIA RTX-GPUs sind auf hohe Skalierbarkeit und Performance für die Analysen und KI-Initiativen von Unternehmen ausgelegt. Dell Technologies liefert umfassende Hardwarelösungen, die für die Ausführung der branchenweit neuesten KI-Software optimiert sind:

- **Robuste Hardwarekonfiguration:** Dell Precision-Workstations punkten durch zahlreiche leistungsstarke Hardwarekonfigurationen, darunter Multi-Core-Prozessoren, RAM mit hoher Kapazität und mehrere GPU-Optionen. Diese Komponenten stellen die benötigten Rechenressourcen für KI-Aufgaben bereit, um effizientes Training und effektive Inferenz zu ermöglichen.
- **Skalierbarkeit und Anpassbarkeit:** Dell Precision-Workstations sind skalierbar und anpassbar, sodass die NutzerInnen die Hardwarekonfiguration auf ihre spezifischen KI-Anforderungen ausrichten können. Diese Flexibilität stellt sicher, dass die Workstation für die besonderen Anforderungen von KI-Workloads optimiert werden kann.
- **Zertifizierung und Optimierung:** Dell arbeitet mit NVIDIA zusammen, um die Kompatibilität und Performance von Precision-Workstations mit NVIDIA RTX-GPUs, einschließlich NVIDIA RTX 6000-Karten der Ada-Generation, zu zertifizieren. Diese Zertifizierung stellt die nahtlose Integration sowie die optimierte Performance bei der Nutzung von Dell Precision-Workstations mit NVIDIA RTX-GPUs für KI-Aufgaben sicher.

- **Leistungsstarke Verarbeitung:** Dell Precision-Workstations mit Intel Prozessoren liefern die für KI-Aufgaben benötigte Rechenleistung. Diese Workstations mit ihren Multi-Core-Prozessoren und hohen Taktraten bieten die Performance, die in KI-Workflows für Training und Inferenz erforderlich ist.
- **Unterstützung durch Software und Tools:** Auf Dell Precision-Workstations sind Software und Tools vorinstalliert, die sowohl die KI-Entwicklung als auch die -Bereitstellung unterstützen. Das umfasst optimierte Software-Stacks, KI-Frameworks sowie Bibliotheken. Diese nutzen die NVIDIA RTX-GPUs und erleichtern den NutzerInnen so die ersten Schritte bei KI-Projekten.

In den nachfolgenden Abschnitten werden weitere Technologien vorgestellt, die ebenfalls wichtige Alleinstellungsmerkmale von Dell Workstations sind.

Reliable Memory Technology

Neben ECC bietet Dell noch eine weitere Technologie mit der Bezeichnung RMT Pro (Reliable Memory Technology Pro) zur Maximierung der Verfügbarkeit. Zusammen mit dem ECC-Arbeitsspeicher lassen sich Arbeitsspeicherfehler in Echtzeit erkennen und beheben. Laut Dell werden Arbeitsspeicherfehler von RTM Pro nahezu eliminiert: Die Technologie verhindert, dass erneut auf den defekten Arbeitsspeicher zugegriffen wird, auch bei voller DIMM-Nutzung. Nach einem Neustart des Systems isoliert RTM Pro den fehlerhaften Arbeitsspeicherbereich und blendet ihn für das BS aus. Folglich tritt bei KI-Data-Scientists und -EntwicklerInnen nicht das Problem auf, dass das System ständig abstürzt, weil noch auf den fehlerhaften Arbeitsspeicher zugegriffen wird. Das führt zu einer enormen Produktivitätssteigerung.

Dell Optimizer for Precision

Dell hat außerdem Dell Optimizer for Precision in die meisten Workstations integriert. Damit werden die Systemeinstellungen automatisch angepasst, sodass etliche gängige kommerzielle Anwendungen mit der höchstmöglichen Geschwindigkeit auf der Workstation ausgeführt werden. Das steigert die Produktivität von Data Scientists und EntwicklerInnen. Zudem erstellt das Tool Echtzeitberichte zur Prozessor-, Storage-, Arbeitsspeicher- und Grafikkartenauslastung für das IT-Team. Bisher kann Dell Optimizer for Precision nicht auf Linux ausgeführt werden. Somit ist das Tool gut für die KI-Bereitstellung geeignet, aber nicht für die KI-Entwicklung, die in der Regel auf Linux-basierter Open-Source-Software erfolgt. Des Weiteren bietet Dell Optimizer for Precision ExpressSign-in, ExpressCharge (für Mobilgeräte), Intelligent Audio sowie Reporting- und Analysetools für das Finetuning der Workstation.

HERAUSFORDERUNGEN/CHANCEN

Für Unternehmen

Laut IDC gibt es eine Spaltung im KI-Markt. Auf der einen Seite führen Unternehmen Datenstrategien ein, um wettbewerbsfähig zu bleiben – darunter auch große KI-Initiativen. Beispielsweise merken sie, dass andere Unternehmen mithilfe von Enterprise-KI-Infrastrukturangeboten, die unter den Top 100 der Supercomputer geführt werden, Außergewöhnliches geleistet haben. Auf der anderen Seite erleben Unternehmen auch die tägliche Realität, in der kleine KI-Initiativen auf den verfügbaren Servern im Rechenzentrum oder in der Cloud getestet werden – häufig mit unzureichendem Budget und leistungsschwacher Hardware.

Für viele Unternehmen ist das erste Szenario nicht relevant und das zweite nur allzu bekannt. Für sie besteht die Herausforderung darin, ihren KI-Data-Scientists und/oder KI-EntwicklerInnen die richtigen Tools zur Verfügung zu stellen, damit das KI-Training in einem angemessenen Zeitraum durchführbar ist, ohne Unsummen für Cloud-Instanzen oder GPU-beschleunigte Rechenzentrumsserver auszugeben. Laut IDC sind diese Unternehmen gut beraten, ihre Data Scientists und EntwicklerInnen mit leistungsstarken GPU-beschleunigten Workstations auszustatten.

Für Dell

Im Markt besteht das Missverständnis, dass für KI-Entwicklung und -Bereitstellung teure, beschleunigte Serverhardware, manchmal sogar ein Cluster erforderlich ist. Das mag für die größten KI-Algorithmen mit Milliarden Parametern zutreffend sein, aber die meisten Unternehmen entwickeln gar nicht solche umfangreichen Algorithmen. Sie möchten vielmehr, dass ihre KI-Initiative nützlich, sinnvoll und verwaltbar ist. Vielen Unternehmen ist gar nicht bewusst, dass sich solch gängige KI-Modelle auf Workstations entwickeln und auch bereitstellen lassen. Die Herausforderung von Dell besteht darin, dieses Missverständnis aufzuklären und den Markt über die Möglichkeiten des Dell Workstation-Portfolios zu informieren.

Gleichzeitig muss Dell sicherstellen, dass die Workstations diese Versprechen auch erfüllen und sich nicht im Laufe der Zeit als technologische Engpässe erweisen. Folglich sind schnelle, kontinuierliche Innovationen nötig, um nie die EndnutzerInnen zu enttäuschen, die diese Workstations ordnungsgemäß einsetzen (die also nicht versuchen, einen Algorithmus mit mehreren Milliarden Parametern auszuführen). Natürlich gibt es für Kunden, die plötzlich sehr schnell skalieren oder deren Algorithmen tatsächlich sehr umfangreich geworden sind, einen nahtlosen Wechsel von der Workstation zur KI-Serverproduktreihe von Dell. Das ist dann wiederum auch eine Verkaufschance für Dell, nämlich jedem Kunden die richtige Lösung anbieten zu können, unabhängig vom Umfang der jeweiligen KI-Initiative.

FAZIT

IDC ist der Ansicht, dass Workstations derzeit als „Zugpferde“ für die KI-Entwicklung und -Bereitstellung in zahlreichen Anwendungsfällen ausgesprochen unterbewertet sind. Sie bieten KI-WissenschaftlerInnen und -EntwicklerInnen eine leistungsstarke GPU-beschleunigte Plattform, die durch niedrigere CAPEX als Server, extrem weniger OPEX als Cloud-Instanzen und deutlich mehr Freiheit für Experimente mit KI-Modellen punktet. Die meisten Unternehmen entwickeln KI-Initiativen, die keine Algorithmen mit Milliarden Parameter erfordern. Sie sollten in Betracht ziehen, ihre KI-Teams mit Workstations auszustatten, die uneingeschränkte KI-Entwicklung und eine einfache Edge-basierte Bereitstellung bieten.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com