ANALYSTENBRIEFING
Gesponsert von Dell

Die Investitionen in Infrastruktur für die KI-Verarbeitung schnellen in die Höhe. Die gute Nachricht ist: Bei über 50 % der Systeme in der KI-Verarbeitungsinfrastruktur ist im Jahr 2024 keine Beschleunigung nötig – sie können mit standardmäßigen Servern und Ethernetnetzwerken ausgeführt werden.

GenAI entwickelt sich in Rekordgeschwindigkeit – Unternehmen starten mit GenAI

Februar 2024

Verfasst von: Brandon Hoff, Research Director, Enabling Technologies: Networking and Comm, und Vijay Bhagavath, Research Vice President, Cloud and Datacenter Networks

Einführung

IDC-Forschung liefert Prognosen und zugrunde liegende Faktoren, die sich unserer Meinung nach im Jahr 2024 und danach auf IT-Investitionen auswirken werden. Technologieführungskräfte und ihre Pendants in den LOBs (Lines of Business) können dieses Dokument als Leitfaden für ihre strategische Planung verwenden.

Bisher haben die Betriebsteams Daten erfasst, Data Lakes erstellt und die Cloud als Daten-Storage genutzt. Dank der zunehmenden Popularität von ChatGPT – dem „iPhone-Moment" für generative KI (kurz: GenAI) – wissen sie nun auch, was sie mit ihrem Datenvolumen machen können. Da die Vorteile von GenAI allgemein bekannt sind, drängen Investoren, Führungskräfte und die Marktlage die Betriebsteams dazu, eine effektive GenAI-Strategie zu implementieren. Es gibt verschiedene Technologien und etliche Optionen, die zur Verbesserung des Geschäftsbetriebs und der Mitarbeiterproduktivität genutzt werden können, von GenAI über ML bis zu digitalen Zwillingen und mehr. Die erfolgreiche Implementierung der am besten geeigneten Technologie wird sich zu einem wesentlichen KPI für Betriebsteams und das Unternehmen insgesamt entwickeln.

AUF EINEN BLICK

DIE WICHTIGSTEN PUNKTE

Beginnen Sie mit der Planung der GenAI-Infrastruktur:

- Integrieren Sie KI schneller als andere, indem Sie zunächst zumindest eine Kopie von Daten On-Premise bereitstellen im Rahmen einer Initiative zur Rückholung von Daten aus der Cloud.
- Sorgen Sie dafür, dass der Wert der Algorithmen, auf denen GenAI, KI, ML und digitale Zwillinge im Business basieren, bekannt ist, und setzen Sie Prioritäten nach geschäftlichem Nutzen.
- Drei Schritte: Stellen Sie standardmäßige Server und Ethernetnetzwerke für die GenAI-Bewertung bereit. Führen Sie das GenAI-Scale-out für Workloads der Enterprise-Klasse ganz nach Bedarf aus (mit Standardethernet). Verteilen Sie Workloads zwischen On-Premise- und externer Infrastruktur, um CAPEX und OPEX in den nächsten 3 bis 5 Jahren zu optimieren.

Momentaufnahme von GenAI

Angesichts der explosionsartig steigenden KI-Nachfrage erweitern Cloud-Serviceanbieter (SPs) und Unternehmen ihre Infrastruktur immer schneller. Cloud-SPs nutzen den Großteil der KI-Beschleuniger und bauen damit ihre eigene KI-Infrastruktur aus. Aber diese Beschleuniger sind teuer und treiben so die Kosten für KI-Services von Cloud-SPs in die Höhe. Zu den KI-Beschleunigern zählen GPUs, TPUs, FPGAs, ASSPs und ASICs. Diese Cloud-SPs bauen quasi „KI-Werke" für massive Workloads, die die Anforderungen von zahlreichen Unternehmen abdecken und in erster Linie in den größten IT-Umgebungen der Welt bereitgestellt werden – also etwa in 9 Konzernen.

Auf der anderen Seite kann die GenAI-Infrastruktur für Workloads der Enterprise-Klasse mit Standardsystemen erstellt werden, und zwar ganz ohne Beschleuniger. Laut IDC-Prognose ist bei über 50 % der Systeme in der KI-Verarbeitungsinfrastruktur im Jahr 2024 keine Beschleunigung nötig. Daher können alle beginnen, ihre GenAI-Infrastruktur mit Standardservern und -netzwerken bereitzustellen. GPUs sind ebenfalls verfügbar, falls sie benötigt werden. Für die Bereitstellung einer KI-Infrastruktur gibt es etliche Optionen – ebenso wie mehrere Typen von GenAI, KI, ML und digitalen Zwillingen, von denen verschiedene Unternehmen auf unterschiedliche Weise profitieren. Die Ausführung von GenAI auf Standardservern bietet Vorteile, da GenAI-Software-Stacks in der Regel unterstützt werden. Unternehmen, die in ihre On-Premise-Standardinfrastruktur investieren, können ihre GenAI-Initiativen daher schneller vorantreiben als andere. Es ist entscheidend, dass IT-Teams damit beginnen, die verschiedenen Algorithmen für GenAI, KI, ML und digitale Zwillinge auszuwerten, um zu ermitteln, welche den größten Nutzen für ihr Unternehmen versprechen.

Vorteile

GenAI und andere grundlegende Modelle sind ein echter Gamechanger: Sie setzen neue Maßstäbe für unterstützende Technologien und bieten auch technisch nicht so versierten NutzerInnen leistungsstarke Funktionen. GenAI hat das Potenzial, Effizienz und Produktivität zu steigern, neue Wachstumschancen zu eröffnen, Kosten zu senken und einen Wettbewerbsvorteil für die Unternehmen zu bieten, die sie nutzen.

Der Aufbau einer eigenen GenAI-Infrastruktur ist der erste Schritt zur Integration dieser bahnbrechenden Technologie in den Geschäftsbetrieb und bringt das Fachwissen vor Ort in den GenAI-Technologiestack ein. Die Priorisierung von Technologieinvestitionen für die Etablierung der initialen GenAI-Infrastruktur On-Premise mit standardmäßigen Enterprise-Servern und Ethernetnetzwerken wird Unternehmen, die diese transformative Technologie nutzen, Vorteile bei der Markteinführung verschaffen.

Überlegungen

Unverzüglicher Einstieg in die GenAI-Nutzung

Der Hype um GenAI basiert auf den beeindruckenden Ergebnissen, die ChatGPT und andere Modelle liefern. GenAI bietet Mehrwert, aber dieser hängt stark von der Quelle der proprietären Daten und den eingesetzten Algorithmen ab. Vorstandsetagen, Investoren und Führungskräfte werden Fragen stellen, um herauszufinden, wie genau GenAI ihrem Unternehmen helfen kann.

Unternehmen, die große Mengen unstrukturierter proprietärer Daten erfasst haben, können mit GenAI Originalinhalte aus diesen proprietären vorhandenen Daten generieren. Mit diesen wiederum kann sich das Unternehmen für kontinuierliche Innovationen neu aufstellen. Der Crawl-Walk-Run-Ansatz ist sinnvoll, um zu bestimmen, was GenAI für das Unternehmen bringen kann und welche Schritte als Nächstes angeraten sind.

Erstellung der ersten GenAI-Infrastruktur mit Ethernet

Für Workloads der Enterprise-Klasse bieten Standardsysteme die nötige Performance für den Einstieg in GenAI. Wenn die GenAI-Infrastruktur auf standardmäßigen Servern und Ethernetnetzwerken basiert, lassen sich zudem Betriebssysteme sowie Management- und Netzwerkmanagementtools der Enterprise-Klasse

verwenden. Zunächst müssen die Compute-Anforderungen der großen Sprachmodelle (Large Language Models, LLMs), von denen die Unternehmen sich Vorteile versprechen, identifiziert werden. Anschließend kann die Optimierung der Compute-Performance mithilfe von geeigneten GenAI- und KI-Beschleunigern erfolgen. Entscheidend ist, dass die Architektur der KI-Infrastruktur gut konzipiert sein muss. Eine Fabric mit guter Architektur kann Dutzende bis Tausende KI-Compute-Nodes unterstützen.

Es gibt verschiedene Netzwerkoptionen für GenAI-Workloads, aber Ethernet wird als universell einsetzbare und offene Lösung von unterschiedlichen Anbietern bevorzugt. Erste GenAI-Bereitstellungen können mit standardmäßigen, bereits verfügbaren Ethernetnetzwerken für GenAI-Cluster unterstützt werden.

Aufbau einer GenAI-Infrastruktur mit Ultra-Ethernet

Da jedes Unternehmen mit seinen eigenen Walk-and-Run-Phasen der GenAI-Entwicklung beginnt, kann es sinnvoll sein, auch eine eigene Scale-out-GenAI-Infrastruktur aufzubauen. Dafür sind zwei weitere wichtige Elemente erforderlich: KI-Beschleuniger im Rechenzentrum und KI-Netzwerke. In einer typischen Scale-out-GenAI-Infrastruktur sind 8 Rechenzentrums-GPUs auf jedem Server vorhanden. Für jede GPU wird eine Hochgeschwindigkeits-NIC oder -DPU bereitgestellt, um Netzwerke mit hoher Performance zu ermöglichen.

„Der Markt der in Rechenzentren für GenAI eingesetzten Ethernetswitches im Enterprise-Segment wird Prognosen zufolge mit einer CAGR von 158,2 % steigen, und zwar von 41,9 Mio. USD im Jahr 2023 auf 1,0 Mrd. USD im Jahr 2027“, Vijay Bhagavath, IDC.

Eine wichtige Voraussetzung für eine Scale-out-KI-Infrastruktur sind hochleistungsfähige Netzwerke. Der Zeitraum, den die Daten im Netzwerk verbringen, ist bei der Verarbeitung von GenAI-LLMs ein Bottleneck. Bei einigen Workloads kann diese Zeitspanne bis zu 60 % der LLM-Verarbeitungszeit betragen – und während die Daten zwischen Compute-Clustern verschoben werden, ist die Compute-Infrastruktur inaktiv. Es gibt schon heute bessere KI-Netzwerke, die vom Ultra Ethernet Consortium bereitgestellt werden. Sie versprechen Verbindungen mit einer Leistungsfähigkeit von Supercomputing-Netzwerken, skalierbar für das Cloud-Rechenzentrum und so kostengünstig sowie universell einsetzbar wie Ethernet. KI-Netzwerke sind unerlässlich, um die wachsenden Netzwerkanforderungen von GenAI und HPC in großem Maßstab zu erfüllen. Die gute Nachricht ist, dass die meisten Anbieter von Ethernetswitches das Ultra Ethernet Consortium unterstützen.

Für die Performance sind drei wichtige Technologiekomponenten erforderlich: Hochgeschwindigkeits-SerDes, PHYs und optische Komponenten. Diese drei Technologien werden in Ethernet- und anderen Netzwerktechnologien verwendet, sodass es im Prinzip bei keiner Netzwerktechnologie einen Leistungsvorteil gibt. Zur Erreichung der höchsten Ethernetperformance hat die InfiniBand Trade Association die RoCE(RDMA over Converged Ethernet)-Initiative ins Leben gerufen und das RoCE-Protokoll definiert. RoCE wird auf standardmäßigen Rechenzentrumsswitches unterstützt. Darüber hinaus gibt es zusätzliche Verbesserungen zur Performancesteigerung, wie z. B. Ethernetswitches mit hoher Radix, Cut-Through-Switching und Lastenausgleich sowie höhere Bandbreiten mit Verbindungen von bis zu 800 GbE (4 x 200 GbE), die in Kürze auf den Markt kommen.

Die initialen Tests von GenAI-LLMs liefern einen frühen Ausblick auf die Vorteile, die GenAI potenziell für Unternehmen bietet. Zudem sind sie hilfreich, um eine Strategie für ein GenAI-LLM zu entwickeln und die benötigten Infrastrukturtypen zu ermitteln. Im Wesentlichen legt der Software-Stack die Halbleiteranforderungen für den nächsten Schritt in der Enterprise-GenAI-Entwicklung fest. Folglich hilft das Verständnis des Software-Stacks bei der Bereitstellung einer optimierten Hardwareinfrastruktur.

Ausgleich von On-Premise- und Off-Premise-Infrastruktur bei stabilisierenden Halbleiterkosten

Mit zunehmender Verfügbarkeit an Rechenzentrums-GPUs werden diese auch von mehr Anbietern verkauft. Mit der Zunahme an KI-Beschleunigern wird eine höhere GenAI-Verarbeitungsgeschwindigkeit für On-Premise-Bereitstellungen zur Verfügung stehen. Parallel dazu werden die Bottlenecks bei den Cloud-Serviceanbietern aufgelöst, daher ist zu erwarten, dass sich die Kosten stabilisieren. Wenn dies geschieht, wird der Ausgleich von GenAI-Workloads zwischen On-Premise- und Cloud-Infrastrukturen in etwa 3 bis 5 Jahren die CAPEX und die OPEX optimieren.

Fazit

GenAI ist die bahnbrechende Technologie für KI. Unternehmen benötigen eine GenAI-Strategie/-Planung, die schon jetzt für Workloads der Enterprise-Klasse umgesetzt werden sollte, um diese bahnbrechende Technologie weiter in den Unternehmensbetrieb zu integrieren.

Die Nachfrage ist hoch und treibt die Preise von Komponenten und Cloud-Serviceanbietern gleichermaßen in die Höhe. Gleichzeitig ist laut IDC-Prognose bei über 50 % der Systeme in der KI-Verarbeitungsinfrastruktur im Jahr 2024 keine Beschleunigung nötig. Daher können alle beginnen, ihre GenAI-Infrastruktur mit Standardservern und -netzwerken bereitzustellen. GPUs sind ebenfalls verfügbar, falls sie benötigt werden. Für die Bereitstellung einer KI-Infrastruktur gibt es etliche Optionen – ebenso wie mehrere Typen von GenAI, KI, ML und digitalen Zwillingen, von denen verschiedene Unternehmen auf unterschiedliche Weise profitieren.

„Der Markt hat das Wachstum von GenAI fortlaufend unterschätzt. IDC erwartet ein stabiles Wachstum bei der GenAI-Infrastruktur und Halbleitern“, Brandon Hoff, IDC.

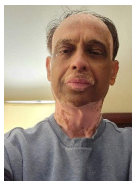
Laut IDC-Prognose werden Unternehmen ihre Daten für die GenAI-Verarbeitung aus der Cloud zurückholen, um die OPEX zu senken. Unternehmen nutzen zunächst standardmäßige Compute- und Ethernetnetzwerkhardware, um GenAI zu entwickeln und zu testen. Wenn sie wissen, welche LLMs für ihr Unternehmen sinnvoll sind und welchen Wert sie aus ihren proprietären Daten ziehen können, folgen weitere Investitionen.

Der Aufbau einer Infrastruktur für GenAI-LLM-Tests mit Standardservern und Enterprise-Ethernetnetzwerken wird den Wert von GenAI für die Unternehmen erschließen.

Informationen zu den AnalystInnen

Brandon Hoff, *Research Director, Enabling Technologies: Networking and Comm*

Brandon Hoff leitet die Netzwerk- und Kommunikationsinfrastruktur von IDC im Enabling-Technologies-Team. Hoff befasst sich mit Technologietrends, Workloads, Produkten, Anbietern, Lieferketten und Strategien für die Akzeptanz von EndnutzerInnen in der Enterprise-IT und den Rechenzentren von Web-, Cloud- und Telekommunikationsserviceanbietern.



Vijay Bhagavath, *Research Vice President, Cloud and Datacenter Networks*

Vijay Bhagavath bietet umsetzbare innovative Konzepte und pragmatische Einblicke in die Märkte und Technologien für Cloud- und Rechenzentrumsnetzwerke. Bhagavath verfügt über umfangreiche Kenntnisse des gesamten Netzwerkmarkts, der Technologien, der Produkt-Roadmaps, der Wettbewerbsdifferenzierung und der Bereitstellungsstrategien, sodass er Providern, Cloud-Anbietern, Enterprise-IT-KäuferInnen und Fachleuten mit aufschlussreichen Erläuterungen und Anleitungen zur Seite stehen kann.

NACHRICHT VOM SPONSOR

IDC Custom Solutions

Der Inhalt dieses Dokuments wurde basierend auf vorhandenen IDC-Studien angepasst, die auf www.idc.com veröffentlicht wurden.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T +1 508 872 8200
F +1 508 935 4015
Twitter @IDC
idc-insights-community.com
www.idc.com