Cool durch den Sommer: Warum Flüssigkeitskühlung das nächste Muss im KI-Bereich ist

Moderne Prozessoren bringen Rechenzentren an ihre thermischen Leistungsgrenzen. Um mithalten zu können, setzen Unternehmen auf Flüssigkeitskühlung, um Leistung, Kosten und Nachhaltigkeit in Einklang zu bringen.

Künstliche Intelligenz ist für einige der anspruchsvollsten Workloads im Unternehmen verantwortlich. Sie diagnostiziert Krankheiten in Echtzeit, optimiert die Logistik in weltweiten Lieferketten und treibt die nächste Welle der Produktinnovation voran. Doch hinter jedem Einblick und hinter jeder Automatisierung steckt ein Server, der auf Hochtouren läuft – und heiß wird.

Die heutigen Hochleistungs-CPUs und -GPUs verbrauchen mehr Strom und erzeugen mehr Wärme als herkömmliche Luftkühlungssysteme jemals bewältigen konnten. Lüfter und Kühlkörper können nicht mithalten, und Rechenzentren zahlen den Preis in Form von Leistungsdrosselung, Energieverschwendung, bedenklichen Lärmpegeln und steigenden Infrastrukturkosten.

Die Flüssigkeitskühlung – einst Supercomputern und Spezialumgebungen vorbehalten – wird heute als praktische Lösung für die thermischen Anforderungen künstlicher Intelligenz immer häufiger eingesetzt. Hier erfahren Sie, wie Direktflüssigkeitskühlsysteme eingesetzt werden, um Wärme effizienter zu verwalten, Energieverschwendung zu verringern und die KI-Infrastruktur im großen Maßstab funktionsfähig zu halten.

Die KI-Hitzekrise

Der Aufstieg der Large Language Models und agentenbasierter KI hat den Bedarf an Rechenleistung in unbekanntes Terrain getrieben. High-End-CPUs verbrauchen mittlerweile regelmäßig mehr als 500 Watt und einige GPUs nähern sich der 1.000-Watt-Marke. Gleichzeitig sinken die Temperaturtoleranzen. Während Chips früher bei 90 bis 100 °C sicher funktionierten, stoßen einige der modernsten Siliziumchips heute bereits bei 60 °C ihre Obergrenze.[1]

Die herkömmliche Luftkühlung ist für diese neue thermische Realität einfach nicht ausgelegt. Um die Wärme moderner Prozessoren zu bewältigen, sind enorme Kühlkörper, leistungsstarke Lüfter und mehr Platz erforderlich, was die Kosten in die Höhe treibt und die Serverdichte reduziert. Es handelt sich um einen nicht effizienten Zyklus, der Leistung und Skalierbarkeit beeinträchtigt.

Außerdem steigt dadurch der Energieverbrauch sprunghaft an. Der Stromverbrauch von Rechenzentren in den USA stieg von 76 Terawattstunden im Jahr 2018 auf 176 TWh im Jahr 2023 und könnte bis 2028 sogar 580 TWh erreichen.[2] Hitze ist zum versteckten Engpass der KI geworden, und steht immer mehr auf dem Spiel.

So funktioniert die Flüssigkeitskühlung

Bei der Flüssigkeitskühlung wird Kühlmittel über die heißesten Komponenten eines Servers – nämlich CPUs und GPUs – zirkuliert, um Wärme aufzunehmen und abzuleiten. Das Kühlmittel zirkuliert durch die an den Hauptkomponenten angebrachten Kühlplatten, leitet die Wärme ab und transportiert sie zu einem externen Verteiler, von wo aus sie zum Wärmetauscher weitergeleitet wird, der die Wärme aus dem System abführt. Das Ergebnis ist eine schnellere und effizientere Wärmeübertragung als mit Luftkühlung möglich ist.

Hewlett Packard Enterprise bietet drei primäre Optionen für die Flüssigkeitskühlung. Closed-Loop-Systeme sind vollständig im Servergehäuse integriert und eignen sich daher ideal für schrittweise Aktualisierungen der vorhandenen Infrastruktur. Die alternative Option ist die Flüssigkeits-Luft-Kühlung, bei der das Betriebswasser zum Kühlen eines an die Racks angeschlossenen Luftkühlungssystems verwendet wird, wodurch kühle Luft gezielter zu den heißesten IT-Bereichen geleitet wird und die Wärme aus der Luft zurück an das Betriebswasser übertragen wird. Schließlich arbeitet die Direktflüssigkeitskühlung auf Rack-Ebene, verteilt Kühlmittel an mehrere Nodes und bietet ein überlegenes Wärmemanagement für dichte, High-Performance-Workloads.

Es gibt auch einen kombinierten Ansatz, bei dem Flüssigkeitskühlung für die stromhungrigsten Komponenten und Luftkühlung für den Rest verwendet wird. Dieses ausgewogene Design stellt sicher, dass jede Komponente effizient gekühlt wird, ohne das System zu überdimensionieren.

Im Vergleich zur Luftkühlung bieten Flüssigkeitssysteme eine bessere Leistung, ermöglichen eine höhere Rackdichte und senken den Stromverbrauch erheblich, wodurch sowohl die Betriebskosten als auch die CO2-Emissionen reduziert werden.

[1] „HPE ProLiant Gen11 Servers with Direct Liquid Cooling", HPE, 2024.

[2] 2024 United States Data Center Energy Usage Report, Lawrence Berkeley National Laboratory, U.S. Department of Energy, Dezember 2024

So setzt HPE Flüssigkeitskühlung im großen Maßstab ein

HPE hat seine Flüssigkeitskühlungslösungen entwickelt, um den Anforderungen moderner KI an Umfang, Dichte und Effizienz gerecht zu werden. Im Mittelpunkt dieser Bemühungen stehen HPE ProLiant Compute Gen12 Server, die Konfigurationen mit Flüssigkeitskühlung unterstützen. Closed-Loop-Systeme bieten eine in sich geschlossene Kühllösung, die sich ideal für CPUs mit hoher Wattzahl in vorhandenen Rack-Setups eignet. Die Direktflüssigkeitskühlung geht noch weiter und verteilt das Kühlmittel über ein ganzes Rack, um Wärme im großen Maßstab zu handhaben – perfekt für dichte, GPU-intensive Workloads.

Für Umgebungen mit gemischten Workloads bietet das Adaptive Cascade Cooling von HPE einen einheitlichen Ansatz. Dieses patentierte System wechselt dynamisch und in Echtzeit zwischen Luft- und Flüssigkeitskühlung, abhängig von der Wärmebelastung der angeschlossenen Komponenten. Das Ergebnis ist eine höhere Kühleffizienz bei niedrigerem Energieverbrauch und geringerer Infrastrukturkomplexität.

Unabhängig davon, welchen Ansatz Sie wählen, die Kosten- und Nachhaltigkeitsgewinne sind erheblich. Die Lösungen können den Energieverbrauch senken, ohne die Leistung zu beeinträchtigen, und sind somit eine strategische Option für Unternehmen, die betriebliche Anforderungen mit Umweltzielen in Einklang bringen müssen. In einer Implementierung mit 10.000 Servern kann die Direktflüssigkeitskühlung von HPE über zwei Millionen US-Dollar an Kühlkosten pro Jahr und gleichzeitig 7,7 Millionen Kilogramm CO_2-Emissionen jährlich einsparen.[3]

Diese Systeme sind bereits auf 100 % lüfterlosen HPE Cray Supercomputern und HPE ProLiant XD Plattformen im Einsatz und unterstützen Bereitstellungen mit hoher Dichte, bei denen herkömmliche Kühlung nicht ausreicht. Da die KI-Workloads weiter zunehmen, bietet die Flüssigkeitskühlung eine praktische Möglichkeit, um den thermischen Anforderungen gerecht zu werden, ohne das gesamte Rechenzentrum neu zu gestalten, da für die Flüssigkeitskühlung keine vollständige Überholung erforderlich ist. Unternehmen können die fortschrittliche Kühlung dort einsetzen, wo sie am dringendsten benötigt wird – beispielsweise bei Workloads mit hoher Dichte oder hoher Hitzeentwicklung. Das ist ein praktischer Schritt zur Erfüllung ökologischer, sozialer und Governance-Ziele.

[3] Liquid Cooling: A Cool Approach for AI, HPE Newsroom Blog, August 2024.

Konzipiert für Wachstum: HPEs Ansatz für eine skalierbare KI

Die innovative Flüssigkeitskühlung von HPE ist nur ein Teil einer umfassenderen Strategie für eine zukunftssichere KI-Infrastruktur. Lösungen wie HPE Private Cloud AI kombinieren flüssigkeitsgekühlte HPE ProLiant Compute Systeme mit NVIDIA-beschleunigtem Computing, integriertem Networking und KI-fähiger Software auf einer sofort nutzbaren Plattform. Diese Systeme wurden für KI im Produktionsmaßstab entwickelt und ermöglichen eine schnellere Time-to-Value bei gleichzeitiger Vereinfachung der Bereitstellung und Skalierung.

Da es sich um lokale Lösungen handelt, behalten Unternehmen die volle Kontrolle über ihre Daten, was für Unternehmen mit Anforderungen an die Datenhoheit, Compliance oder geringe Latenz von Vorteil ist. Gleichzeitig erleichtern die vorhersehbare Wirtschaftlichkeit und integrierte, effiziente Kühlung das Management der Betriebskosten bei steigenden KI-Anforderungen.

HPE AI Services kann Unternehmen bei der Planung und Implementierung einer Infrastruktur unterstützen, die den Leistungs- und auch den thermischen Anforderungen von KI-Workloads gerecht wird. Dazu gehört die Auswahl des richtigen Ansatzes für die Kühlung, der zum Bereitstellungsumfang und der Art der Workloads passt. Diese Art der Abstimmung – zwischen Rechenleistung, Kühlung und Betrieb – kann die Komplexität verringern und eine effiziente Erweiterung im Laufe der Zeit erleichtern.

Der Hitze immer einen Schritt voraus

Die Entwicklung der KI wird nicht langsamer und auch die Anforderungen, die sie an die Infrastruktur stellt, werden anspruchsvoller. Da die Chips immer heißer und die Workloads immer komplexer werden, kann die herkömmliche Kühlung nicht mehr Schritt halten. Durch die Flüssigkeitskühlung ändert sich die Art und Weise, wie Unternehmen ihre Infrastrukturplanung angehen müssen. Durch effizienteres Wärmemanagement unterstützt sie Bereitstellungen mit höherer Dichte, senkt den Stromverbrauch und trägt zur Verlängerung der Hardwarelebensdauer bei – alles unerlässlich für die nachhaltige Skalierung von KI-Workloads.

In der Zukunft der KI geht es heiß her – aber dank der Flüssigkeitskühlung kann wenigstens Ihre Infrastruktur cool bleiben.

Weitere Informationen finden Sie unter

HPE.com/ai/insights

HPE.com besuchen

Jetzt chatten

a50013353DEE

HEWLETT PACKARD ENTERPRISE

hpe.com