Die Cloud in der Maschine – 19-Zoll-Technik als skalierbares Datenrückgrat

Warum ein über hundert Jahre alter Standard immer noch top-aktuell ist

Blick in das Rechenzentrum von CERN © Wikipedia / Florian Hirzinger

In modernen Industriemaschinen und Fertigungsstraßen sind zahlreiche Kameras im Einsatz, deren Bilder möglichst in Echtzeit ausgewertet werden sollen. Aufgrund der hohen Daten-Bandbreite und der geringen geforderten Latenz entwickelt sich die Cloud-Anbindung zu einem Flaschenhals. Abhilfe schafft es, die Cloud in die Maschine hineinzuverlagern. Hierfür eigenen sich 19-Zoll-Rechnereinschübe. EFCO montiert diese auf der Grundlage eines selbst entwickelten Baukastensystems jetzt in Deutschland, um kurze Lieferzeiten bei skalierbarer Auslegung zu gewährleisten.

Die 19-Zoll-Rechner für High-End-Computing von EFCO tragen unterschiedlichen Anforderungen Rechnung. Liegt der Fokus weniger auf Datendurchsatz, sondern mehr auf Datenverwaltung und den Einsatz trainierter KI-Modelle, dann sind lüfterlose Ausführungen möglich. Die 3-HE-Einschübe sind dabei lediglich 270 mm tief und erlauben die Montage in dezentralen kleinen Rechnerschränken direkt an der Maschine (Bild 1).

Autor:
Tom Weber
Robologs
EFCO Electronics GmbH
www.efcotec.de

Hi-Performance-Computing ohne Lüfter?

Wie in Bild 1 sichtbar, sind die Seitenflächen der IPC-Gehäuse als zusätzliche, konventionelle Kühlkörper ausgeführt. Damit diese ihre Wirkung entfalten können, ist eine entsprechende Thermik im Rack erforderlich. Diese kann auf natürlichem Weg erfolgen (warme Luft steigt auf), indem z. B. gekühlte Luft (z. B. aus der Klimaanlage) am Boden eingeblasen und in Deckenhöhe wieder abgesaugt wird. Wird ein höher Luftdurchsatz benötigt, bieten sich Lüfter-Einschübe an, die unten im Rack montiert werden, und die Luft nach oben, an den Einschüben vorbei, zum Strömen zu bringen.

Zwangsgekühlte Varianten

Werden die anfallenden Verlustleistungen zu groß, um mit Hilfe natürlicher Konvektion abgeführt zu werden, kommen zwangsgekühlte Varianten zum Einsatz. Bei EFCO sind deren Mainboards entweder für die Generationen 12 mit 14 von Intel optimiert - oder für die Xeon-6-Prozessoren desselben Herstellers (Bild 2). Die neue Generation der Rack-IPCs von EFCO umfasst vier Formfaktoren von 1 bis 4 HE (Bild 3). Allen Geräten gemeinsam ist die Ausstattung mit den üblichen Schnittstellen von Analog über USB und HDMI bis zu 2,5 GbE.

Ebenso zeitgemäß ist die Security-Ausstattung mit TPM 2.0. Einsetzen kann man diese Rechner bei Umgebungstemperaturen von 0…50 °C. Um dies zu ermöglichen, verwendet etwa EFCO möglichst große, temperaturgesteuerte Lüfter mit zweifacher Kugellagerung für jahrzehntelange Laufruhe. Je nach Gehäuseformat sind bis zu sieben PCI- bzw. PCIe-Erweiterungssteckplätze vorhanden, bis zu vier Montageplätze für hot-swap-fähige 2,5-Zoll-SATA-Datenspeicher sowie optische Laufwerke (Bild 4).

Skalierbare KI-Beschleuniger-Hardware

Gerade für das Training von KI wird – genauso wie für anspruchsvolle Bild- und Videoverarbeitung – häufig eine dedizierte KI-Hardware erforderlich. An dieser Stelle bündeln die beiden taiwanesischen Unternehmen EFCO und aetina ihre Kräfte. Die ausgeklügelten KI-Beschleuniger der umfangreichen Jetson-Serie von aetina bleiben extern und lassen sich über eine oder mehrere 2.5-GbE-Schnittstellen problemlos an den IPC anbinden. In dieser Kombination stehen bis zu 275 TOPS an Rechenpower bereit; dafür sind aber auch bis zu 130 W Leistung erforderlich.

Rack-IPC mit Nvidia-Karte

Für den späteren Einsatz vor Ort: Der limitierende Faktor für die "on-board"-KI-Hardware ist meist die maximal verfügbare interne Versorgungsleistung; häufig stehen nicht mehr als 65 W zur Verfügung. Dies reicht aber völlig aus, um etwa eine Einsteckkarte mit einem Nvidia Orin-NX (mit bis zu 100 TOPS) oder einem Nvidia Orin-Nano mit bis zu 40 TOPS zu versorgen. Die KI-Beschleuniger-Hardware ist damit nicht mehr extern, sondern im Gehäuse des Rack-IPCs gut geschützt und direkt an das Mainboard angebunden.

Rack-IPC mit M.2-Einsteckkarte

Die kleine Lösung: Alle Motherboards der Rack-IPCs von EFCO verfügen über einen M.2-Erweiterungsport, der eigentlich für SSDs gedacht ist. Da diese aber auch gut über die schnellen USB-3.0-Schnittstellen angebunden werden können, lässt sich der M.2-Slot als PCI-Express-Port nutzen. Auch hier ist die für den Steckerverbinder spezifizierte maximale elektrische Leistung der limitierende Faktor. Dennoch sind mit KI-Beschleunigern als M.2-Einsteckkarte bis zu 25 TOPS möglich.

SaaS-Backend

Ein weiterer Einsatzbereich der hochperformanten Rack-IPCs ist die Nutzung von SaaS-Geschäftsmodellen. Hierbei wird die eigentliche Software nicht auf dem Zielsystem des Kunden zur Verfügung

Bild 1: Kompakter Rack-IPC von EFCO mit konventioneller Kühlung (lüfterlos).

Bild 2: Die kompakten 1-HE-Rack-IPCs sind mit knapp 400 mm Gehäusetiefe sehr kompakt.

gestellt, sondern online als Dienstleistung (SaaS: Software as a Service). Damit dieses System mit dem potenziellen Flaschenhals „Uplink-Verbindung zum Server“ seine Vorteile ausspielen kann, ist eine leistungsfähige Vorverarbeitung der Daten ebenso erforderlich, wie die Umsetzung und Berechnung der zurückgelieferten Ergebnisse und Parameter.

Von Vorteil für SaaS ist grundsätzlich eine High-Speed-Uplink-Verbindung zum Server. Diese wird heute häufig über 2,5 GbE hergestellt – oder per Glasfaser. Entsprechend sollten zeitgemäße Rack-IPCs die Möglichkeit bieten, SFP-Module einzusetzen.

Bild 3: Die kompakten äußeren Abmessungen erfordern eine durchdachte „Inneneinrichtung“.

Bild 4: In den 2-HE-Rack-IPCs findet bereits ein optisches Laufwerk Platz. Das Gehäuse ist 508 mm tief.

Im High-End-Computing wird Entwärmung immer wichtiger

Wo hohe Rechenleistungen gefordert sind, fällt in aller Regel auch entsprechende Verlustwärme an. Während man sich früher um die abzuführenden Wärmemengen eher wenig Gedanken gemacht, und die thermischen Verluste „einfach“ mit Ventilatoren (im allgemeinen als „Zwangskühlung“ umschreiben) in die Raumluft weggekühlt hat, werden die Anforderungen zunehmend komplexer. Einerseits bedeutet ein unkontrolliertes Aufheizen der Umgebungsluft, dass diese Luft dann z. B. über Klimaanlagen wieder heruntergekühlt werden muss. Dies ist ein ineffizientes Verfahren mit hohem Stromverbrauch.

So entfallen in großen Rechenzentren beispielsweise bis zu 40 % der eingesetzten Energie auf das Wärmemanagement. Entsprechend geht der Trend zu flüssig gekühlten Einheiten. Dafür spricht nicht nur, dass Wasser in Verhältnis zu Luft die 4.000-fache Wärmemenge aufnehmen und damit abtransportieren kann, sondern dass die Wärme im Wasser technisch einfacher nutzbar ist – von der Heizung bis hin zur Ab(d)sorbitions-Kältemaschinen. Es gibt also funktionierende Lösungen, um aus (Ab-)Wärme (nutzbare) Kälte zu erzeugen.

Ein einfaches Rechenbeispiel

Ein High-End-Rechner benötigt maximal 300 m³/h Kühlluft. Zehn solcher Rechner befinden sich in einem Rack. Dann müssen im Extremfall pro Stunde 3.000 m³ Luft durch das Rack bewegt werden. Damit dies überhaupt gelingt, ist sind entsprechende konstruktive Aufwände erforderlich (z. B. Luftleitbleche); zum anderen erreicht diese Form der Kühlung in großen Rechenzentren die Grenzen des physikalisch Möglichen; die Lüfter sind einfach nicht mehr in der Lage, genügend Luft durch die Einschübe zu bewegen, weil die anderen Lüfter im Schrank genau das Gleiche tun – was zu gegenseitigen Staudrücken führt. Oder andersherum: Zwischen der Ansaugseite und der Ausblasseite eines Lüfters muss ein gewisser Druckunterschied bestehen, damit die Luft durch den Ventilator strömen kann und nicht durch die anderen Lüfter zurückgedrückt wird.

Statt der 3.000 m³ Luft im Beispiel oben, könnte man die gleiche Wärme auch mit Hilfe von 750 l Wasser pro Stunde abführen – dafür reicht eine einfache Heizungspumpe für unter 100 Euro.

19-Zoll-Praxistipp

Beim Kauf von Netzwerkschränken immer auf die Abmessungen (vor allem die Tiefe!) der Geräte achten, in die die Racks eingebaut werden sollen. Rack-Hersteller geben in aller Regel die Außenbreite und -tiefe an; lediglich die verfügbare Einbauhöhe wird in HE angegeben. Ein Schrank mit der Angabe 600 x 800 mm, 42 HE hat entsprechend eine Außenbreite von 600 mm sowie eine Außentiefe von 800 mm und bietet Platz für Geräte mit in Summe 42 HE (ca. 1,87 m). Die nutzbare Innenbreite (z. B. für die Kabelführung seitlich an den Einschüben vorbei) sowie die nutzbare Innentiefe (wichtig z. B. für leistungsfähige Rack-IPCs) sind von Hersteller zu Hersteller unterschiedlich! ◄

19-Zoll – die Basics

Die 19-Zoll-Technik wurde 1922 durch das American Standards Institute (ANSI) etabliert. In Rechenzentren wie auch in der Veranstaltungstechnik wird der Begriff „Rack“ häufig für ein Gestell (oder einen Schaltschrank) verstanden, in welches Geräte mit einer Breite von 19 Zoll (= 48,26 cm) eingeschoben werden – daher der Begriff „Einschub“. In der Labortechnik werden häufig „leere“ Einschübe verwendet, die dann als Baugruppenträger bezeichnet werden. Diese verfügen über Schienen, damit sich kleinere Einheiten in der 19-Zoll-Technik montieren lassen.

Die Höhe aller Einschübe wird in Höheneinheiten (HE) angegeben, wobei eine Höheneinheit 1,75 Zoll oder 44,45 mm entspricht. Ein klassisches Format ist 3 HE – also eine Gerätehöhe von 133,35 mm. Die Breite wird in aller Regel in Teilungseinheiten (TE) angegeben, wobei eine Teilungseinheit 1/5 Zoll und damit 5,08 mm entspricht. Schmale 19-Zoll-Geräte für Baugruppenträger (häufig auch Einschubkassetten genannt) sind üblicherweise 12 TE breit – entsprechend also 60,96 mm. Damit es wegen unvermeidlicher Maßtoleranzen keine Probleme gibt, spezifiziert der Standard eine negative Toleranz (Mindermaß) von maximal 1/32 Zoll (= 0,79 mm) pro Frontplatte. Entsprechend beträgt die reale Breite einer 12-TE-Einschubkassette also zwischen 60,17 und 60,96 mm. Fischer elektronik etwa gibt 60,4 mm an.

Die Mechanik der 19-Zoll-Technik ist international genormt (EIA 310-D, IEC 60297 sowie DIN 41494 SC48D) und erleichtert damit die Kompatibilität von Geräten unterschiedlichster Hersteller bei der Montage in einem 19-Zoll-Rack.

Bild 5: Eine herausziehbare 19-Zoll-Rackkonsole – Monitor und Tastatur in einem 1-HE-Einschub. Die Flansche zur Montage im 19-Zoll-Rack sind unter dem Monitor (vorne) sowie hinten an der Teleskop-Schiene erkennbar. © Fokus Technologies GmbH