DELTA COMPUTER

PRODUKT-DOSSIER · DGX STATION

NVIDIA DGX Station.

Frontier-KI am Schreibtisch. Grace-Blackwell-Ultra-Superchip, 748 GB kohärenter Speicher, 20 PFLOPS NVFP4: produktionsreif ab Auslieferung. Drei OEM-Varianten. Eine Architektur.

748 GB	20 PFLOPS	1.600 W	3 OEM
KOHÄRENTER SPEICHER	NVFP4-INFERENZ	230-V-STECKDOSE	VARIANTEN

01 · EINSTIEG

Frontier-KI ohne Rechenzentrum.

Die DGX Station ist NVIDIAs erstes Frontier-KI-System für den Arbeitsplatz. Sie kombiniert den Grace-Blackwell-Ultra-Superchip mit 748 GB kohärentem Speicher und 20 PFLOPS NVFP4 in einer Workstation, die an einer normalen 230-V-Bürosteckdose läuft. Modelle, die zuvor einen Cluster erforderten, werden lokal trainiert, fein-tuned und betrieben, ohne Cloud, ohne Wartelisten.

KI am Schreibtisch

Geschlossen flüssigkeitsgekühlt, 1.600 W an 230 V, lautlos im Bürobetrieb. Kein Serverraum, kein zusätzlicher Wasseranschluss, keine RZ-Klimatisierung. Vom ersten Tag produktionsreif.

GB300 Superchip

Grace-72-Core-Arm-CPU und Blackwell-Ultra-GPU als ein gemeinsamer Adressraum. 900 GB/s NVLink-C2C zwischen CPU und GPU. 252 GB HBM3e direkt am Beschleuniger.

Datenhoheit

Trainingsdaten, Modellgewichte und Inferenz bleiben auf dem Gerät. Keine Cloud-Übertragung, keine Drittstaaten, keine Wartelisten. On-Prem-Betrieb passt zum DSGVO- und Schrems-II-Rahmen.

Cluster-VM oder lokale Station?

CLOUD-VM (TYPISCH)

Geteilte Hardware, geteilte Wartezeit

- Zugriff auf ungekannte Hardware-Topologien
- Trainingsdaten verlassen den Standort
- Laufende Stundenkosten, Verfügbarkeit nicht garantiert
- Compliance-Argumentation gegenüber Datenschutz schwer

DGX STATION LOKAL

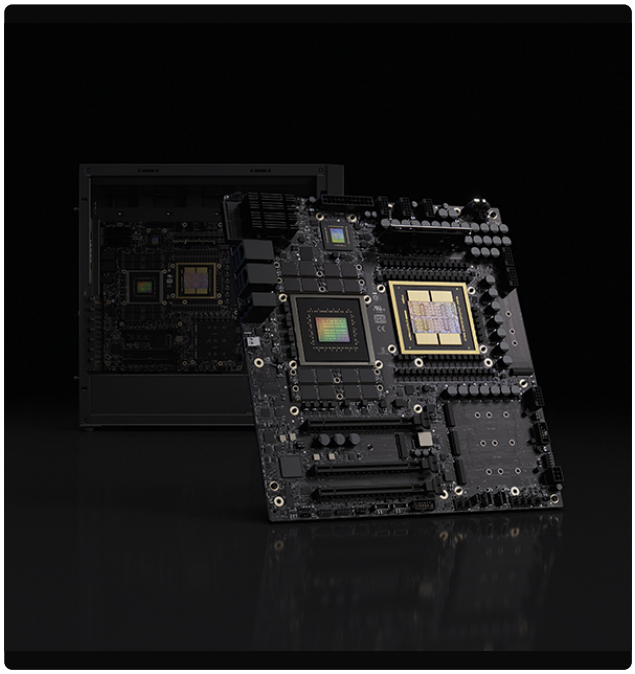
Eigene Hardware, planbare Kosten

- Bekannte Topologie, identisch zum SuperPOD
- Daten und Modelle bleiben lokal
- Einmalige Investition, klare Lebenszyklus-Kosten
- Compliance-fähig nach DSGVO und Schrems II

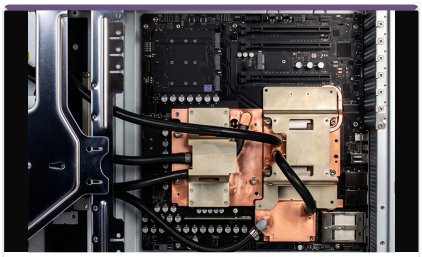
02 · ARCHITEKTUR

Grace und Blackwell Ultra als ein System.

Der GB300 Desktop Superchip integriert die klassischen CPU-/GPU-Steckkarten der Workstation in einen einzigen, eng gekoppelten Beschleuniger, verbunden über 900 GB/s NVLink-C2C zu einem einheitlichen Adressraum von 748 GB. Keine PCIe-Engpässe, keine expliziten Daten-Kopien zwischen Host und Device.



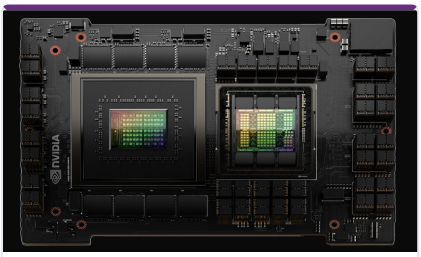
Drei Schlüsselkomponenten



KÜHLUNG

Geschlossener Kühlkreislauf

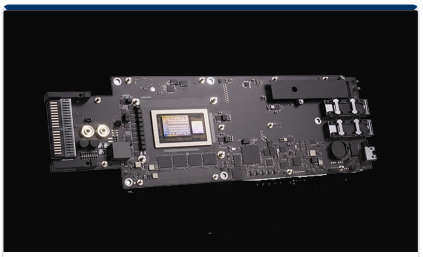
Closed-Loop Flüssigkeitskühlung für leisen, bürotauglichen Betrieb bei 1.600 W Systemleistung.



RECHENEINHEIT

GB300 Superchip

Grace-CPU und Blackwell-Ultra-GPU kohärent verbunden über NVLink-C2C mit 900 GB/s.



NETZWERK

ConnectX-8 Networking

Zwei QSFP-Ports mit je 400 GbE. Zwei Stations koppelbar ohne externen Switch.

NVFP4	3,5× kompakter als FP16. Natives Format für Frontier-Modelle, das ein 670-Mrd.-Modell überhaupt erst auf eine einzelne Station passen lässt.
Grace CPU	72-Core Arm Neoverse V2 direkt im System. Trägt Datenpipeline, Tokenisierung und Pre-/Post-Processing ohne separate x86-Host-CPU.

03 · FORMFAKTOR

Ein Schreibtisch. Eine Steckdose. Frontier-KI.

Die DGX Station bringt die Architektur eines Frontier-KI-Rechenzentrums an einen einzigen Arbeitsplatz. Was ein 230-V-Anschluss, ein Ethernet-Kabel und vier m² Stellfläche ermöglichen, hat zuvor einen 19-Zoll-Schrank, dedizierte Kühlung und einen Rechenzentrums-Vertrag erfordert.

748 GB	20 PF	1.600 W	7
Modellspeicher	**NVFP4 sparse**	**Anschlussleistung**	**MIG-Instanzen**
Kohärent · CPU+GPU	15 PFLOPS dense	230-V-Bürosteckdose	Multi-User auf einer GPU

Was darauf läuft

FP8-Inferenz typisch bis ~400 Mrd. dense bzw. 671 Mrd. MoE. Volles Fine-Tuning unter BF16 auf 70-Mrd.-Modellen in HBM. NVFP4-Inferenz bis in die Billion-Parameter-Klasse mit LPDDR-Pool.

Was übrig bleibt

Bekannter CUDA-Stack, identisch zur HGX-, NVL72- und SuperPOD-Linie. Code, Container und Pipelines lassen sich ohne Änderungen vom Schreibtisch in den Cluster verschieben.

QUELLE

Werte nach NVIDIA DGX Station (GB300) Spec-Tabelle, Stand April 2026 [1]. Fine-Tuning- und Inferenz-Werte als Richtgrößen der CUDA-NVFP4-Toolchain [2].

04 · SPEZIFIKATIONEN

Technische Daten im Überblick.

Identische Hardware-Performance über alle drei OEM-Varianten. Unterschiede betreffen nur Formfaktor, Netzteilklasse und Betriebstemperaturbereich (siehe Seite 6).

SUPERCHIP & RECHENEINHEIT

Architektur	**Grace Blackwell Ultra**
Superchip	**1× NVIDIA GB300 Desktop**
GPU	**20.480 CUDA Cores · NVFP4**
CPU	**1× Grace · 72 Arm Neoverse V2**

SPEICHER

GPU-Speicher	**252 GB HBM3e · 7,1 TB/s**
CPU-Speicher	**496 GB LPDDR5X · 396 GB/s**
Gesamt (kohärent)	**748 GB einheitlich**
NVLink-C2C	**900 GB/s CPU ↔ GPU**

KI-PERFORMANCE

NVFP4	**20 PFLOPS sparse**
NVFP4 dense	**15 PFLOPS**
FP8 / FP6	**10 PFLOPS**
FP16 / BF16	**5 PFLOPS**
Multi-Instance GPU	**Bis zu 7 isolierte Instanzen**

NETZWERK & ERWEITERUNG

Netzwerk	**ConnectX-8 SuperNIC · 800 Gb/s**
Management	**1× RJ45 10 GbE · 1× 1 GbE BMC**
PCIe-Slots	**1× Gen5 ×16 · 2× Gen5 ×16 (×8 elektr.)**
Storage	**4× M.2 Gen5 NVMe**

System & Betrieb

ANSCHLUSSLEISTUNG

Eingangsleistung	**1.600 W · 230 V**

KÜHLUNG

System	**Closed-Loop Liquid Cooling**

BETRIEBSSYSTEM

OS	**NVIDIA DGX OS 7 · Ubuntu 24.04**

05 · OEM-VARIANTEN

Drei Ausführungen, identische Performance.

DELTA liefert die DGX Station in drei OEM-Ausführungen auf Basis des identischen NVIDIA-GB300-Referenzdesigns. Die Wahl betrifft Formfaktor, Designsprache und Service-Ökosystem, nicht die Rechenleistung.

GIGABYTE

W775-V10-L01

SUPERMICRO

ARS-511GD-NB-LCC

MSI

XpertStation WS300

Merkmal	GIGABYTE	SUPERMICRO	MSI
Formfaktor	Tower	Tower / Rack	Tower
Abmessungen B×H×T	248 × 528 × 716 mm	455 × 218 × 781 mm	248 × 528 × 583 mm
Netzteil	1.600 W Platinum	1.600 W Titanium	1.600 W Titanium
Betriebstemperatur	10–35 °C	10–25 °C	10–35 °C
Eingangsspannung	230 V (Schuko)	230 V (Schuko)	230 V (Schuko)
IDENTISCH ÜBER ALLE DREI VARIANTEN			
Superchip	**1 × NVIDIA GB300 Desktop Superchip**		
Kohärenter Speicher	**748 GB (252 GB HBM3e + 496 GB LPDDR5X)**		
NVFP4-Performance	**20 PFLOPS sparse · 15 PFLOPS dense**		
Netzwerk	**ConnectX-8 SuperNIC · 800 Gb/s**		
Kühlung	**Closed-Loop Liquid Cooling**		
Betriebssystem	**NVIDIA DGX OS 7 · Ubuntu 24.04**		

06 · SKALIERUNG & SOFTWARE

Ein Stack.
Vom Schreibtisch zum SuperPOD.

Der CUDA-Stack ist über die Spark-/Station-/B300-/NVL72-/SuperPOD-Linie identisch. Code, Container und Pipelines lassen sich vom Arbeitsplatz in den Cluster verschieben, ohne dass die Anwendung neu validiert werden muss.

Skalierungsleiter

ENTRY	SIE SIND HIER	SERVER	RACK	MULTI-RACK
DGX Spark	**DGX Station**	**DGX B300**	**GB300 NVL72**	**DGX SuperPOD**
GB10 128 GB · 1 PFLOPS 170 W	**GB300** 748 GB · 20 PFLOPS 1.600 W	**8× Blackwell Ultra** 2,3 TB · 72 PFLOPS ~15 kW	**72 × GB300** ~21 TB HBMe 1,1 EFLOPS FP4	**576 GPUs** pro Scalable Unit Multi-EFLOPS

NVIDIA-AI-Software-Stack

ENTERPRISE

NVIDIA AI Enterprise

Validierte Container, Frameworks und Microservices. 90-Tage-NVAIE-Lizenz im Lieferumfang.

FRAMEWORK

NeMo · NIM · RAPIDS

NeMo für LLM-Training und Fine-Tuning, NIM für lokale Inferenz-Microservices, RAPIDS für Datenpipeline.

INFERENZ

TensorRT-LLM · vLLM · SGLang

Validierte Open-Source-Serving-Stacks, sofort lauffähig auf der DGX OS-Basis.

BETRIEBSSYSTEM

NVIDIA DGX OS 7

NVIDIA-validiertes Ubuntu-24.04-Image mit kompletten Treibern und Kubernetes-Werkzeugen.

ÖKOSYSTEM

Hugging Face · PyTorch · vLLM · JAX

Volle Kompatibilität mit dem gängigen Open-Source-Stack. Trainings- und Inferenz-Code aus der Forschungs-Community läuft ohne Anpassung.

VERWALTUNG

Mission Control · Run:ai

Cluster-weite Orchestrierung und GPU-Scheduling, sobald der Schritt vom Schreibtisch zum Cluster ansteht.

07 · ANWENDUNGSFELDER

Was passt in 748 GB?

Der einheitliche Adressraum aus HBM3e und LPDDR5X verändert die Frage, was lokal noch handhabbar ist. Die folgenden vier Profile beschreiben, wo die DGX Station typischerweise zum Einsatz kommt: Inferenz bis in die Billion-Parameter-Klasse.

FORSCHUNG & HOCHSCHULEN

LLM-Forschung, Reinforcement Learning, multimodale Modelle

Lokales Pre-Training auf 70-Mrd.-Klasse, Fine-Tuning bis in den dreistelligen Milliarden-Bereich, MoE-Inferenz darüber hinaus. DFG-fähige Großgeräte-Beschaffung mit klarer Hardware-Liste, Slurm-tauglich.

LIFE SCIENCES & PHARMA

BioNeMo, MONAI, Wirkstoff-Modelle

Lokales Fine-Tuning auf Patienten- und Wirkstoffdaten, ohne Übertragung in eine Cloud. FP64-fähige Beschleunigung für Molekülsimulation. Kompatibel mit dem Compliance-Rahmen klinischer Forschungseinrichtungen.

INDUSTRIE & ENGINEERING

Simulation, CAE-/CAD-AI, Robotik (GR00T)

Lokale agentische Robotik-Workflows, FEM-/CFD-AI-Surrogate, generatives Engineering. Trainingsdaten und Hersteller-IP bleiben innerhalb des Standorts, ohne Cloud-Übertragung in Drittstaaten.

SOFTWARE-ENTWICKLUNG & PROTOTYPING

Coding-Assistenten, Agenten-Pipelines, Inferenz-Server

Bis zu sieben isolierte MIG-Instanzen für Multi-User-Entwicklung. Prototypisches Inferenz-Serving direkt am Arbeitsplatz, mit klarem Skalierungspfad in den Cluster, sobald die Workload trägt.

08 · DELTA & DGX STATION

Validiert von NVIDIA. Integriert von DELTA.

DELTA Computer Products ist NVIDIA Elite Partner und 3× Star Performer Central Europe. Wir liefern, installieren und begleiten DGX-Systeme über den gesamten Lebenszyklus, vom Einzelplatz bis zum Frontier-SuperPOD.

DELTA TEST-CENTRE · GLINDE

Hardware vor der Auslieferung prüfen.

Im hauseigenen Test-Centre integriert und validiert DELTA jede DGX Station vor der Auslieferung. Burn-in-Tests, Performance-Validierung und Konfigurations-Setup laufen vor dem Versand, sodass das System am ersten Tag am Standort produktionsreif läuft.

Referenz aus DELTA-Projekten

DELTA-INTEGRATION · 2025

DeepL „Arion“

DGX GB200 NVL72 SuperPOD

DGX GB200 NVL72 SuperPOD für DeepL, von DELTA in Europa als einer der ersten integriert. Direkter Skalierungspfad vom Einzelplatz auf der DGX Station bis in die NVL72-Klasse. Identische CUDA-Toolchain, identisches Software-Bild.

Was DELTA für die DGX Station leistet

Beratung
Workload & OEM-Auswahl

Lieferung
4–10 Wochen, DACH-weit

Installation
Aufbau am Standort

Schulung
Admins & Anwender

Wartung
OEM + DELTA-Service

Finanzierung
Leasing über Bankpartner

ISO 9001
Qualitätsmanagement

ISO 14001
Umweltmanagement

ISO 27001
Informationssicherheit

EcoVadis Gold
Top 5 % Nachhaltigkeit

09 · BESCHAFFUNG

Vom Erstgespräch zur produktiven Station.

Trotz Workstation-Formfaktor bleibt die DGX Station eine NVIDIA-validierte Frontier-Plattform. Die folgenden fünf Schritte beschreiben den typischen Beschaffungsweg für DACH-Kunden in Forschung und Industrie.

1

Bedarfsklärung

Workload-Profil, Modellgrößen, Aufstellungsort und Beschaffungsweg im Erstgespräch.

2

Site-Survey

230 V-Anschluss, Stellfläche, Netzwerkanbindung und Lautstärke prüfen.

3

OEM & Angebot

MSI, Supermicro oder Gigabyte auswählen. Brutto-Angebot für die Beschaffungsstelle.

4

Lieferung & Aufbau

4–10 Wochen ab Bestellung. Inbetriebnahme und Performance-Test vor Ort.

5

Betrieb

Wartung, Schulung und Erweiterung in den Cluster über den gesamten Lebenszyklus.

Compliance-Rahmen

EU-Hoheit

Lieferung, Montage und Service über DELTA als deutsches Unternehmen. Kein US-Vertragspartner in der Lieferkette.

Schrems II

On-Prem-Argumentation: Trainingsdaten, Modellgewichte und Inferenz bleiben innerhalb der DACH-Region.

DSGVO

Volle Kontrolle über Trainingsdaten und Modellgewichte. Lokale Verarbeitung ohne Cloud-Übertragung.

Export

US Advanced Computing Chips Rule und EU-Dual-Use-Verordnung 2021/821: ausschließlich an gewerbliche Kunden bzw. Körperschaften des öffentlichen Rechts.

DGX STATION · LIEFERBAR AB MAI 2026

Eine Station.
Eine Steckdose.
Ein Ansprechpartner.

Wir beraten Sie zur passenden OEM-Variante, prüfen die Voraussetzungen am Standort und begleiten die Beschaffung bis zur produktiven Inbetriebnahme.

UNTERNEHMEN

DELTA Computer
Products GmbH
Am Alten Lokschuppen 4
D-21509 Glinde

KONTAKT

Tel +49 40 300672-0
info@delta.de
deltacomputer.com

ÖFFNUNGSZEITEN

Mo–Fr 07:30–18:30
Außerhalb der Öffnungszeiten
nach Vereinbarung

DELTA COMPUTER

10 · QUELLEN & HINWEISE

Belege und Rechtliche Hinweise.

[1] **NVIDIA DGX Station (GB300). Specifications.** 252 GB HBM3e, 496 GB LPDDR5X, 748 GB einheitlich, 20 PFLOPS NVFP4 sparse, 1.600 W an 230 V, Closed-Loop Liquid Cooling.
www.nvidia.com/en-us/data-center/dgx-station/

[2] **NVIDIA Technical Blog & CUDA-NVFP4-Toolchain.** NVFP4-Inferenz- und Fine-Tuning-Werte als Richtgrößen für Foundation-Modelle in den Klassen 70–671 Mrd. Parameter.
developer.nvidia.com/blog · NVIDIA NeMo & TensorRT-LLM

[3] **OEM-Datenblätter.** MSI XpertStation WS300 (CT60-S8060), Supermicro Super AI Station (ARS-511GD-NB-LCC), Gigabyte W775-V10-L01. Alle auf identischer GB300-Referenz-Plattform.
msi.com · supermicro.com · gigabyte.com

> NVIDIA-Disclaimer wörtlich: „Specifications are subject to change without notice." Maßgeblich für jede konkrete Angebot- und Vertragsbasis ist das zum Bestellzeitpunkt gültige NVIDIA- bzw. OEM-Datenblatt.

MARKEN UND BEZEICHNUNGEN

KUNDENREFERENZEN

Die in diesem Dokument genannte Integration des DGX GB200 NVL72 SuperPOD „Arion" für DeepL (2025) wurde von DELTA Computer Products GmbH umgesetzt.

LIEFER- UND VERTRAGSBEDINGUNGEN

DELTA Computer Products GmbH liefert ausschließlich an gewerbliche Kunden und Körperschaften des Öffentlichen Rechts. Für die hier gezeigten DGX-Station-Konfigurationen gelten je nach Konfiguration und Empfängerland die US-Exportkontrollvorschriften (Advanced Computing Chips Rule) sowie die EU-Dual-Use-Verordnung 2021/821; die deutsche Außenwirtschaftsverordnung (AWV) gilt ergänzend. Es gelten ausschließlich unsere AGB (deltacomputer.com/agb). Preise und Verfügbarkeiten freibleibend, vorbehaltlich Zwischenverkauf.

UNTERNEHMENSDATEN

DELTA Computer Products GmbH · Am Alten Lokschuppen 4 · D-21509 Glinde · Deutschland · Tel +49 40 300672-0 · Geschäftsführer: Hans-Peter Hellmann · Amtsgericht Lübeck, HRB 3678 RE · USt-IdNr. DE135110550