DELTA COMPUTER

— PRODUKT-DOSSIER · VERA RUBIN PLATTFORM

NVIDIA Vera Rubin.

Die nächste Generation der NVIDIA Rack-Scale-Architektur.
Sieben co-designte Chips, NVL72-Rack, 3,6 EFLOPS NVFP4-Inferenz.
Verfügbar ab H2 2026.

7	3,6 EF	75 TB	H2 26
CO-DESIGNTE CHIPS	NVFP4-INFERENZ	FAST MEMORY	VERFÜGBAR

— 01 · EINSTIEG

Eine Fabrik für künstliche Intelligenz.

Vera Rubin ist die neue Generation der NVIDIA-Plattform für KI-Infrastruktur. Sie wurde gebaut, um große Sprach- und Reasoning-Modelle effizient zu trainieren und zu betreiben. Die Plattform wird in spezialisierten Racks geliefert, die NVIDIA gemeinsam mit über 80 Hardware-Partnern fertigt. Ein einzelnes Rack der NVL72-Klasse kombiniert die Rechenleistung von 72 Hochleistungs-GPUs zu einem zusammenhängenden System.

Was ist eine AI Factory?

Eine AI Factory ist ein Rechenzentrum, das nicht nur Daten verarbeitet, sondern KI-Modelle als Endprodukt erzeugt. Vergleichbar mit einer Produktionsanlage. Vera Rubin ist die Hardware-Basis, auf der solche Modelle entstehen, trainiert und in Echtzeit beantwortet werden.

Was ist Vera Rubin konkret?

Vera Rubin ist eine Plattform aus sieben aufeinander abgestimmten Chips: Prozessoren (Vera CPU), Beschleuniger (Rubin GPU), Verbindungs-Chips (NVLink, ConnectX, BlueField, Spectrum) und ein dediziertes Inferenz-System (Groq 3 LPX). Diese Chips sind so konstruiert, dass sie als ein einziges System zusammenarbeiten.

Warum eine eigene Plattform?

Eigene Vera-Rubin-Systeme im Rechenzentrum sichern Datenhoheit, Vertraulichkeit und planbare Kosten. Drei Anforderungen, die in Forschung, Industrie und regulierten Branchen zentral sind. Trainingsdaten und Modellgewichte verlassen das eigene Haus nicht.

Der Schritt von Blackwell zu Vera Rubin

BISHER · BLACKWELL / GB300

Eingeführt 2024–2025

- HBM3e-Speicher
- NVLink 5 Scale-up
- 800-Gbit/s ConnectX-8
- Bis zu 1,4 EFLOPS FP4-Inferenz pro NVL72-Rack

→

NEU · VERA RUBIN / RUBIN ULTRA

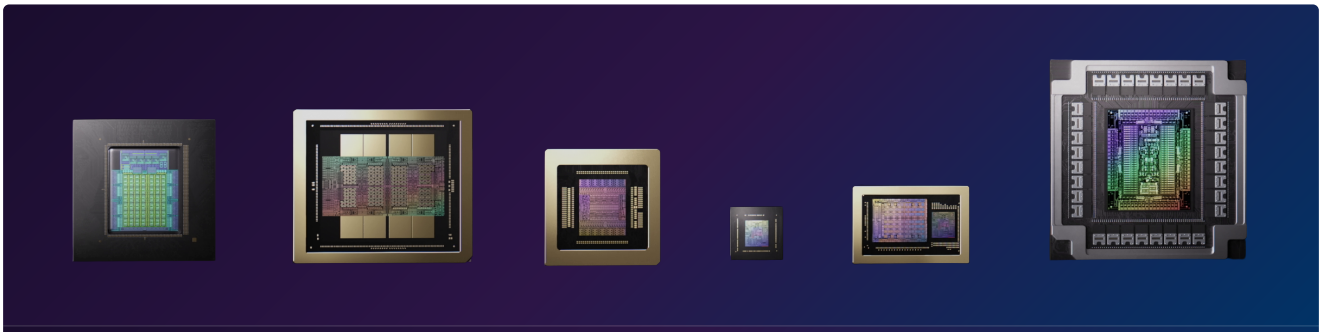
Verfügbar ab H2 2026

- HBM4-Speicher (288 GB pro GPU)
- NVLink 6 mit doppelter Bandbreite
- 1,6-Tbit/s-Klasse ConnectX-9
- 3,6 EFLOPS NVFP4-Inferenz pro NVL72-Rack

02 · ARCHITEKTUR

Sieben Chips. Ein System.

Vera Rubin ist die erste NVIDIA-Plattform, in der Compute, Memory, Scale-up- und Scale-out-Netzwerk sowie Inferenz-Beschleunigung als **geschlossenes System** co-designt sind. Engpässe entstehen dadurch nicht zwischen einzelnen Komponenten, sondern werden im gesamten System aufgelöst.



Vera-Rubin-Chip-Familie.

01
Vera CPU
Arm-basiert, Olympus-Kerne (Armv9.2). 88 Kerne und 176 Threads pro Sockel.

02
Rubin GPU
288 GB HBM4 pro GPU, 22 TB/s Speicherbandbreite. 50 PFLOPS NVFP4.

03
NVLink 6 Switch
3,6 TB/s pro GPU, 200G SerDes. Doppelte Bandbreite gegenüber Blackwell.

04
ConnectX-9 SuperNIC
1,6-Tbit/s-Klasse, InfiniBand und Ethernet, Scale-out pro GPU.

05
BlueField-4 DPU
Daten-, Storage- und Sicherheitsdienste auf jedem Compute-Tray. ICMS für KV-Cache-Offloading.

06
Spectrum-6 Switch
Ethernet-Switching für Cluster und SuperPOD. Optional Quantum-X800 InfiniBand als Scale-out-Variante.

07
NVIDIA Groq 3 LPX
Dedizierte Inferenz-LPU für disaggregiertes Serving im Verbund mit Rubin.

7

Chips, eine geschlossene Plattform. Compute, Memory, Scale-up und Scale-out. Co-designed.

— 03 · LEISTUNGSSPRUNG

Mehr Leistung. Weniger Watt. Weniger Kosten.

Laut NVIDIA: bis zu 5× Inferenzleistung, 3,5× Trainingsleistung, 10× geringere Token-Kosten und 4× weniger GPUs für MoE-Training gegenüber Blackwell. Vergleichswerte basieren auf von NVIDIA gewählten Benchmark-Szenarien.

5×
Inferenzleistung
NVFP4-Peak pro GPU [1]

3,5×
Trainingsleistung
NVFP4-Peak pro GPU [1]

10×
geringere Token-Kosten
Rack-Vergleich Reasoning [2]

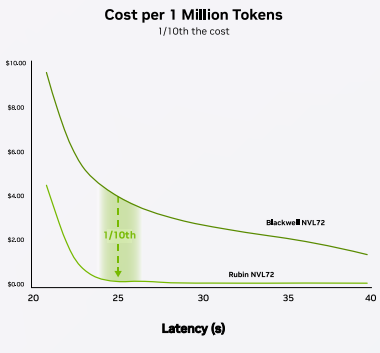
4×
weniger GPUs
für MoE-Training [2]

4× weniger GPUs

10T-MoE-Modell, 100T Tokens, 1 Monat. NVIDIA [2]

1/10 Token-Kosten

Kimi-K2-Thinking, 32K/8K. NVIDIA [2]



QUELLE

Werte nach NVIDIA NVFP4-Peak, NVIDIA Newsroom Stand April 2026 [1] [2]. Vergleich gegen Blackwell-Referenzkonfigurationen.

04 · RACK-SCALE

NVL72 im Detail.

Das NVL72-Rack ist die Basisplattform der Generation: 72 Rubin GPUs und 36 Vera CPUs in 18 Compute-Trays. Erstes NVIDIA-Rack-Scale-System mit kabelloser, schlauchloser, lüfterloser Bauweise und 100 % Direct-Liquid-Cooling.

COMPUTE

GPUs	**72 × Rubin GPU**
CPUs	**36 × Vera CPU**
CPU-Kerne	**3.168 Olympus-Kerne**
Compute-Trays	**18 pro Rack**

MEMORY

HBM4 gesamt	**20,7 TB**
HBM4 pro GPU	**288 GB**
Bandbreite	**1.580 TB/s**
Fast Memory gesamt	**75 TB**

PERFORMANCE

NVFP4 Inferenz	**3.600 PFLOPS**
NVFP4 Training	**2.520 PFLOPS**
FP8 Training	**1.260 PFLOPS**
FP16 / BF16	**288 PFLOPS**

VERBINDUNG

NVLink 6 gesamt	**260 TB/s**
NVLink pro GPU	**3,6 TB/s**
NVLink-C2C	**1,8 TB/s**
Scale-out NIC	**ConnectX-9**

KÜHLUNG & STROM

Kühlung	**100 % liquid**
Inlet-Temperatur	**bis 45 °C**
Bauweise	**kabel-/schlauch-/lüfterlos**
Rack-Busbar	**5.000 A liquid-cooled**

ZUVERLÄSSIGKEIT

RAS	**2. Generation**
In-System-Tests	**Zero-Downtime**
SRAM-Repair	**In-Field**
Tray-Wechsel	**Hot-Swap**

> Vera Rubin NVL72 nutzt die MGX-Oberon-Rack-Architektur mit 5.000-A-Liquid-Busbar. Direct-Liquid-Cooling auf Rack-Ebene ist Voraussetzung; DELTA prüft die Standortvoraussetzungen vor dem Angebot.

— 05 · NVLINK

Doppelte Bandbreite gegenüber Blackwell.

NVLink 6 verdoppelt die Pro-GPU-Bandbreite von Blackwell auf 3,6 TB/s und vervierfacht sie gegenüber Hopper. Lane-Speed steigt von 100G auf 200G PAM4 bei identischer Anzahl von 18 Links pro GPU laut NVIDIA-GTC-Materialien.

Generation	Version	Bandbreite pro GPU	Links pro GPU	Lane-Speed	Plattform
Hopper	NVLink 4	900 GB/s	18	100G PAM4	H100 / H200
Blackwell Ultra	NVLink 5	1,8 TB/s	18	100G PAM4	GB200 / GB300
Rubin	**NVLink 6**	**3,6 TB/s**	18	**200G PAM4**	**Vera Rubin NVL72**

Was NVLink 6 ermöglicht

Mit 260 TB/s Scale-up-Bandbreite pro Rack ist NVL72 effektiv ein Memory-Pool über 72 GPUs. Modelle, die zuvor zwischen Knoten verteilt werden mussten, laufen in einer einzigen NVLink-Domäne — ohne PCIe-Engpass, ohne RDMA-Overhead. Das vereinfacht Tensor-Parallelismus und ermöglicht bei MoE-Modellen Expert-Parallelism in Hardware-Geschwindigkeit.

> SerDes-Details (18 Lanes × 200G PAM4) gemäß NVIDIA-GTC-Materialien. Die 3,6 TB/s pro GPU sind die Spec-Tabelle auf nvidia.com/vera-rubin-nvl72/.

06 · PORTFOLIO

Fünf Varianten. Ein Baukasten.

Vom 8-GPU-Server bis zur AI Factory: Die Vera Rubin Plattform deckt das gesamte Skalierungsspektrum ab. Von der schlüsselfertigen DGX-Variante über das HGX-Baseboard bis zum DGX SuperPOD als Scalable-Unit-Modell.

RACK · TURNKEY — 72 GPUs

NVIDIA DGX™ Vera Rubin NVL72

Schlüsselfertige, NVIDIA-validierte Variante des NVL72-Racks. Inklusive Mission Control, AI Enterprise, DGX OS und 3 Jahren Business-Standard-Support [4].

3,6 EF	75 TB	3 J.
NVFP4	FAST MEMORY	SUPPORT

Auch als OEM-Variante (Supermicro, Gigabyte, ASUS u.a.) verfügbar.

8-GPU · TURNKEY · X86 — 8 GPUs

NVIDIA DGX™ Rubin NVL8

8 Rubin GPUs mit 2 × Intel Xeon 6 Host-CPUs für x86-Continuity. Direct-Liquid-Cooling, validierte Turnkey-Variante.

2,3 TB	400 PF	Xeon 6
HBM4	NVFP4	HOST

8-GPU · BASEBOARD — 8 GPUs

NVIDIA HGX™ Rubin NVL8

OEM-Baseboard mit 8 Rubin GPUs, Wahl zwischen Vera CPU oder x86. Bei Supermicro, Gigabyte, ASUS, Dell, HPE u. a.

2,3 TB	400 PF	5,5×
HBM4	NVFP4	VS. HGX B200

INFERENZ-RACK · LPU — 256 LPUs

NVIDIA Groq 3 LPX

Der siebte Chip: 256 LPUs pro Rack, 128 GB SRAM, 40 PB/s On-Chip-Bandbreite. Disaggregiertes Serving via NVIDIA Dynamo. Co-designed mit Vera Rubin NVL72.

128 GB	40 PB/s	35×
SRAM	SRAM-BW	PER WATT [5]

SUPERPOD · AI FACTORY — 576 GPUs

NVIDIA DGX SuperPOD™ mit Vera Rubin

Scalable-Unit-Modell: 8 Racks, 576 Rubin GPUs pro Scalable Unit. Mehrere SUs kombinierbar bis 60 EFLOPS (Vera Rubin POD, 40 Racks). Vollständig flüssigkeitsgekühlt.

28,8 EF	166 TB	60 EF
NVFP4 / SU	HBM4 / SU	POD MAX

07 · VERGLEICHSMATRIX

Welche Variante für welchen Einsatz?

Die folgende Matrix fasst die zentralen technischen Eckwerte aller fünf Vera-Rubin-Produkte zusammen.

✓ unterstützt / verfügbar — nicht zutreffend für diese Variante

Kategorie	DGX Vera Rubin NVL72	DGX Rubin NVL8	HGX Rubin NVL8	Groq 3 LPX	DGX SuperPOD
Formfaktor	Rack (turnkey)	8-GPU Turnkey	8-GPU Baseboard	Inferenz-Rack	Multi-Rack
GPU-Anzahl	72	8	8	—	576
Host-CPU	36 × Vera	2 × Xeon 6	Vera oder x86	x86 + FPGA	288 × Vera
Speicher	20,7 TB HBM4	2,3 TB HBM4	2,3 TB HBM4	128 GB SRAM	166 TB HBM4
Inferenz-Peak	3,6 EF NVFP4	400 PF NVFP4	400 PF NVFP4	315 PF FP8	28,8 EF NVFP4
NVLink-Domäne	72 GPUs	8 GPUs	8 GPUs	—	72 GPUs + IB
x86-fähig	—	✓	✓	✓	—
Kühlung	100 % liquid	liquid (DLC)	liquid + L2A	liquid	100 % liquid
Verfügbarkeit	H2 2026	H2 2026	H2 2026	H2 2026	H2 2026

Die DGX Vera Rubin NVL72 ist die schlüsselfertige Variante des NVL72-Racks. Identische Architektur und Spezifikation auch als OEM-Variante (Supermicro, Gigabyte, ASUS u.a.) verfügbar [3].

— 08 · SOFTWARE

Vollständiger NVIDIA-AI-Stack.

Die DGX-Varianten und der DGX SuperPOD enthalten den vollständigen NVIDIA-Software-Stack. HGX- und Vera-Rubin-NVL72-Plattformen erlauben wahlweise NVIDIA AI Enterprise oder den eigenen, gewohnten Stack.

VERWALTUNG
NVIDIA Mission Control
Cluster-weite Orchestrierung, Fehlermanagement, Monitoring.

ENTERPRISE-SOFTWARE
NVIDIA AI Enterprise
Validierte Container, Frameworks, Microservices mit Enterprise-Support.

SCHEDULING
NVIDIA Run:ai
Workload- und GPU-Scheduling für gemeinsame KI-Cluster.

BETRIEBSSYSTEM
NVIDIA DGX OS
Auf Ubuntu basierend, NVIDIA-validiertes OS-Image.

ORCHESTRIERUNG
Slurm und Kubernetes
Klassisches HPC-Scheduling und cloud-native Orchestrierung.

FRAMEWORK
NVIDIA NeMo
NeMo Run und Megatron Core für Modell-Training und Fine-Tuning.

INFERENZ
NVIDIA Dynamo
Disaggregiertes Inferenz-Serving für Reasoning-Workloads.

INFERENZ-FRAMEWORKS
TensorRT-LLM, vLLM, SGLang
Validierte Open-Source-Serving-Stacks.

DATA / SECURITY
NVIDIA DOCA
DPU-Programmiermodell für BlueField-Workloads, mit ICMS / Inference Context Memory Storage.

09 · EINSATZSZENARIEN

Welche Variante für welchen Kunden?

Die folgende Empfehlung ist als Orientierung gedacht. Jede Variante wird konkret geplant: DELTA prüft Workload-Profil, vorhandene Infrastruktur und Beschaffungsweg gemeinsam mit Ihnen.

FORSCHUNG & HOCHSCHULEN

DGX Vera Rubin NVL72 oder HGX Rubin NVL8 mit Vera CPU

FP64, konvergente HPC-/AI-Workloads, Slurm, DFG-Antragsbegleitung. Für Drittmittel-finanzierte Großgeräte (DFG-Antrag, Art. 91b GG): turnkey, validiert, mit klarem Wartungskonzept und vollständiger Hardware-Liste für die Begründung der Geräteklasse.

AUTOMOTIVE & INDUSTRIE

DGX SuperPOD für Training; Groq 3 LPX für Sensor-Inferenz

Großformatiges Training für Omniverse, Cosmos und AV-Modelle. Im Verbund mit Inferenz-Racks für nahezu Echtzeit-Sensor-Verarbeitung. On-Prem-Betrieb hält Trainingsdaten und Hersteller-IP innerhalb des eigenen Standorts, ohne Cloud-Übertragung.

LIFE SCIENCES & PHARMA

DGX Vera Rubin NVL72 für lokales Training auf sensitiven Datensätzen

FP64-Genauigkeit für Wirkstoff- und Molekülsimulation, BioNeMo-Stack für proteinbezogene Modelle, lokales Fine-Tuning ohne Übertragung von Patientendaten oder Wirkstoff-Pipelines. On-Prem-Betrieb passt zum Compliance-Rahmen von Pharma- und klinischen Forschungseinrichtungen.

GROSSINDUSTRIE & MANUFACTURING

HGX Rubin NVL8 (mit x86) oder DGX Rubin NVL8

x86-Continuity für bestehende Workflows. RL-Engine, agentische Robotik (GR00T), Kompatibilität mit etablierter Linux-Infrastruktur. Klare Skalierungspfade vom 8-GPU-System bis zum Multi-Rack-Cluster bei wachsendem Bedarf.

CLOUD- & SERVICE-PROVIDER

DGX Vera Rubin NVL72 + Groq 3 LPX im Verbund

Disaggregiertes Serving für niedrige Token-Kosten und hohe Token-pro-Watt-Effizienz. Laut NVIDIA bis zu **1/10 der Token-Kosten** gegenüber Blackwell NVL72 (Kimi-K2-Thinking, 32K Input / 8K Output) [2]. Ein zentrales Argument im AI-Service-Markt.

— 10 · BESCHAFFUNG

Vorbereitung, Beschaffung, Inbetriebnahme.

Vera-Rubin-Systeme erfordern aufgrund ihrer Leistungsdichte sowohl technische als auch organisatorische Vorbereitung. Die folgenden fünf Schritte beschreiben den typischen Beschaffungsweg für DACH-Kunden im Forschungs- und Industrie-Umfeld.

Beratung in deutscher Sprache

Vera-Rubin-Spezialisten klären Workload und Sizing vor dem Angebot.

Vor-Ort-Installation durch eigene Techniker

DACH-weite Inbetriebnahme, kein Subunternehmer in der Lieferkette.

Zertifiziert nach ISO 9001 / 14001 / 27001

Qualitätsmanagement, Umwelt, Informationssicherheit.

1 **Bedarfsklärung**

Workload-Profil, Modellgrößen, Datenhoheit, Standortvorgaben.

2 **Standort-Check**

Stromzuführung, Wasserkreislauf, Raumstatik, Inlet-Temperatur.

3 **Variante & Angebot**

Konfiguration, Brutto-Angebot für die Beschaffungsstelle.

4 **Lieferung & Installation**

Vor-Ort-Aufbau, NVIDIA-validierte Inbetriebnahme, Performance-Test.

5 **Betrieb & Wartung**

Lifecycle-Begleitung, Hardware-Support, Erweiterungen.

Compliance-Rahmen

EU-Hoheit

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

Schrems II

On-Prem-Argumentation: Daten und Modelle bleiben innerhalb der DACH-Region. Kein Datenaustausch in Drittstaaten.

DSGVO

Volle Kontrolle über Trainingsdaten und Modellgewichte. On-Prem-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.

— 11 · DELTA & VERA RUBIN

Validiert von NVIDIA. Integriert von DELTA.

DELTA Computer Products ist seit 2018 NVIDIA Elite Partner und wurde mehrfach als NVIDIA Star Performer Central Europe ausgezeichnet. Wir liefern, integrieren und begleiten Vera-Rubin-Systeme über den gesamten Lebenszyklus.

Referenz aus DELTA-Projekten

DELTA-INTEGRATION · 2025

DeepL „Arion"

DGX GB200 NVL72 SuperPOD

GB200-NVL72-SuperPOD für DeepL, von DELTA integriert. Direkter Skalierungspfad für den Übergang auf die Vera-Rubin-Generation.

Was DELTA für Vera Rubin leistet

Beratung
Workload & Sizing

Konfiguration
Hardware & Stack

Lieferung
DACH-weit

Installation
Aufbau & Inbetriebnahme

Schulung
Admins & Anwender

Wartung
Support & SLA

Finanzierung
Leasing & Bank-Partner

Compliance
Export & DSGVO

— VERA RUBIN · 2. HALBJAHR 2026

Sieben Chips. Eine AI-Fabrik. Ein Ansprechpartner.

Wir beraten Sie zur passenden Variante, prüfen die Voraussetzungen am Standort und begleiten die Beschaffung bis zur 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

— 12 · GLOSSAR

Begriffe auf einen Blick.

Kompakte Erklärung der Fachbegriffe und Abkürzungen, die in diesem Dossier verwendet werden.

NVFP4

Vier-Bit-Gleitkomma-Format von NVIDIA für Inferenz-Workloads. Reduziert Speicherbedarf und Bandbreite gegenüber FP8/FP16, ohne Modellqualität signifikant zu verlieren.

NVLink 6 / NVLink-C2C

NVIDIAs Hochgeschwindigkeits-Verbindung zwischen GPUs (NVLink) bzw. zwischen CPU und GPU (Chip-to-Chip). Bei Vera Rubin: 3,6 TB/s pro GPU bzw. 1,8 TB/s pro Superchip.

MGX

NVIDIAs modulare Referenz-Architektur für Rack-Scale-Designs, die von OEM-Partnern (Supermicro, Gigabyte, ASUS u.a.) implementiert wird.

HGX

GPU-Baseboard, das von OEM-Partnern in eigene Server-Plattformen integriert wird. Im Gegensatz zu DGX kein fest definiertes Komplettsystem.

MoE / Mixture of Experts

Neuronale Netzwerk-Architektur, bei der nur ein Teil der Modellparameter pro Inferenz aktiv ist. Skaliert auf sehr große Modelle bei moderaten Compute-Kosten.

DOCA / ICMS

DPU-Programmiermodell für BlueField-Workloads (Data Center-on-a-Chip Architecture). ICMS = Inference Context Memory Storage für KV-Cache-Offloading.

LPU

Language Processing Unit. NVIDIA Groq 3 LPX ist ein dediziertes Inferenz-Rack mit 256 LPUs und 128 GB SRAM, optimiert für niedrige Latenz.

HBM4

High Bandwidth Memory der vierten Generation. Direkt auf der GPU-Trägerplatine verbauter Speicher mit deutlich höherer Bandbreite als klassischer DRAM.

NVL72 / NVL8

Rack-Scale-Konfiguration mit 72 GPUs (NVL72) bzw. 8-GPU-Baseboard (NVL8). Beide bilden eine geschlossene NVLink-Domäne.

DGX

Schlüsselfertige, NVIDIA-validierte System-Variante mit Mission Control, AI Enterprise und Business-Standard-Support.

SuperPOD

Multi-Rack-Cluster aus mehreren NVL72-Racks. Mindesteinheit „Scalable Unit" mit 8 Racks (576 GPUs).

ConnectX-9 / BlueField-4

NVIDIAs SuperNIC (1,6-Tbit/s-Klasse) für Scale-out-Netzwerk und DPU für Daten-, Storage- und Sicherheitsdienste auf jedem Compute-Tray.

PAM4

Pulse Amplitude Modulation, 4 Levels. Modulationsverfahren der NVLink-/InfiniBand-Lanes mit 200 Gbit/s pro Lane bei NVLink 6.

DFG-Großgerät

Beschaffung wissenschaftlicher Großgeräte in Forschungseinrichtungen über Drittmittel der Deutschen Forschungsgemeinschaft (DFG-Antrag, Art. 91b GG).

— 13 · QUELLEN

Belege und Datenherkunft.

Alle in diesem Dossier mit [N] markierten technischen Aussagen verweisen auf die folgenden offiziellen NVIDIA-Quellen, geprüft zum Stichtag 28. April 2026.

[1] **NVIDIA Vera Rubin NVL72 — Specifications.** Per-GPU NVFP4-Peak-Compute (50 PFLOPS Inferenz, 35 PFLOPS Training) gegenüber Blackwell-GPU.
www.nvidia.com/en-us/data-center/vera-rubin-nvl72/
NVIDIA Technical Blog: „Inside the NVIDIA Vera Rubin Platform". developer.nvidia.com

[2] **NVIDIA Newsroom (CES 2026, 5. Januar 2026; GTC 2026, 16. März 2026).** Performance-Vergleiche zu Blackwell-Konfigurationen: 4× weniger GPUs für MoE-Training (10T-MoE-Modell, 100T Tokens, 1 Monat Trainingsfenster); 1/10 Token-Kosten (Kimi-K2-Thinking, 32K Input / 8K Output Sequenz-Längen).
nvidianews.nvidia.com/news/rubin-platform-ai-supercomputer
nvidianews.nvidia.com/news/nvidia-vera-rubin-platform

[3] **NVIDIA MGX-Partnerprogramm.** Vera Rubin NVL72 Rack-Architektur ist Teil der 3. Generation MGX und verfügbar bei mehr als 80 NVIDIA-MGX-Partnern weltweit, darunter Supermicro, Gigabyte, ASUS, Dell, HPE, QCT.
www.nvidia.com/en-us/data-center/products/mgx/

[4] **NVIDIA DGX Vera Rubin NVL72 — Produktseite.** Schlüsselfertige NVIDIA-validierte Variante mit NVIDIA Mission Control, AI Enterprise, DGX OS. Genaue Support-Laufzeiten und Service-Bundles werden über das NVIDIA-Partnerportal abgewickelt.
www.nvidia.com/en-us/data-center/dgx-vera-rubin-nvl72/

[5] **NVIDIA Vera Rubin NVL72 — Spec-Tabelle (Groq 3 LPX co-design).** Wörtliches NVIDIA-Zitat: „co-designed with Vera Rubin NVL72 to deliver 35× inference performance per watt and up to 10× more revenue opportunity for trillion parameter models relative to Blackwell."
www.nvidia.com/en-us/data-center/vera-rubin-nvl72/

NVIDIA-Disclaimer wörtlich: „Preliminary information. All values are up to and subject to change. Projected performance subject to change." Maßgeblich für jede Angebot- und Vertragsbasis ist das zum Bestellzeitpunkt gültige NVIDIA-Datenblatt.

— 14 · HINWEISE

Rechtliche Hinweise.

QUELLEN UND VALIDITÄT

Die in diesem Dokument enthaltenen technischen Werte stammen aus den offiziellen NVIDIA-Quellen zum Stichtag 28. April 2026: nvidia.com/en-us/data-center/vera-rubin-nvl72/, nvidia.com/en-us/data-center/dgx-vera-rubin-nvl72/, nvidia.com/en-us/data-center/dgx-rubin-nvl8/, nvidia.com/en-us/data-center/hgx/, NVIDIA Newsroom-Mitteilungen vom 5. Januar 2026 (CES) und 16. März 2026 (GTC) sowie NVIDIA Technical Blog (developer.nvidia.com). NVIDIA kennzeichnet diese Werte ausdrücklich als ***„Preliminary information. All values are up to and subject to change. Projected performance subject to change."*** Maßgeblich für jede konkrete Angebot- und Vertragsbasis ist das zum Bestellzeitpunkt gültige NVIDIA-Datenblatt.

MARKEN UND BEZEICHNUNGEN

NVIDIA, das NVIDIA-Logo, NVIDIA DGX, NVIDIA HGX, NVIDIA DGX SuperPOD, NVIDIA Vera Rubin, NVIDIA Rubin, NVIDIA Blackwell, NVIDIA Hopper, NVIDIA Mission Control, NVIDIA AI Enterprise, NVIDIA Run:ai, NVIDIA NeMo, NVIDIA Dynamo, NVIDIA NVLink, NVIDIA ConnectX, NVIDIA BlueField, NVIDIA Spectrum, NVIDIA Quantum, NVIDIA MGX, NVIDIA Groq und NVIDIA Olympus sind eingetragene Marken oder Marken der NVIDIA Corporation. Andere genannte Marken und Eigennamen der jeweiligen Inhaber.

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 Vera-Rubin-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. Lieferungen außerhalb der DACH-Region bitte vorab anfragen. Es gelten ausschließlich unsere Allgemeinen Geschäftsbedingungen (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 · StNr. 11 30 292 31888 57