

NVIDIA Deep Learning Server

Leistungsbeschreibung

1 Inhaltsverzeichnis

1	INHALTSVERZEICHNIS	2
2	LEISTUNGSBESCHREIBUNG	3
2.1	Leistungsgegenstand	3
2.2	Lieferung von Hardware	4
2.3	Serviceleistungen	4
2.4	Gewährleistung und Garantie	4
2.5	Verbleib aller Datenspeicher beim Auftraggeber	5
2.6	Entsorgung	5
3	ANHÄNGE	6
3.1	Anhang 1: Serverkomponente	6
3.2	Anhang 3 NVIDIA DGX Systems Enterprise Services	7

2 Leistungsbeschreibung

Im Folgenden werden die durch den künftigen Auftragnehmer zu erbringenden Liefer- und Dienstleistungen beschrieben, die mit dieser Ausschreibung beschafft werden.

2.1 Leistungsgegenstand

Gegenstand dieser Ausschreibung ist die Beschaffung eines NVIDIA-Blackwell-basierten Deep Learning Servers **NVIDIA DGX B300**. Es wird für die in **Anhang 1 (Deep-Learning-Server)** beschriebene Hardware, deren Einzelteile und deren Service und Support ausgeschrieben.

2.1.1 Grundkonfiguration

Der Deep Learning Server ist in **Anhang 1** beschrieben und enthält 8 NVIDIA-Blackwell DGX B300 Grafikprozessoren (GPU) mit jeweils 288 GB GPU-Speicher, somit 2304 GB GPU-Speicher insgesamt. Die Grafikkarten haben eine CUDA Compute Capability mindestens auf Version 9 und sind über NVLink miteinander verbunden. Die Bandbreite zwischen zwei beliebigen GPUs muss mindestens 1800 GB/s bidirektional betragen, die aggregierte GPU-Interkonnektivitätsbandbreite pro Server 14 TB/s (non-blocking). Die GPU Speicherbandbreite muss mindestens 7.0 TB/s pro GPU betragen. Die Rechenoperationen auf den Grafikkarten müssen mindestens 72 petaFLOPS, gemessen mit FP8-Datentyp und Sparse-Matrix-Option, sowie mindestens 144 petaFLOPS, gemessen mit FP4 Tensor Core Datentyp und Sparse-Matrix-Option betragen.

Ein Server wird mit 2 TB DDR5-basiertem Arbeitsspeicher ausgeliefert, der auf 4 TB pro Server erweiterbar ist ohne, dass die bereits verbauten DDR5-Module ersetzt werden müssen. Ein Server ist mit mindestens zwei x86/x64 kompatiblen Prozessoren mit jeweils mindestens 64 physikalischen Kernen (128 Threads) pro Prozessor ausgestattet. Die Prozessoren unterstützen den PCIe 5.0 Standard und es sind mindestens 88 Lanes pro Prozessor verfügbar.

Der Server verfügt über acht OSFP Ports für NVidia ConnectX-8 mit bis zu 800 Gb/s Infiniband Konnektivität sowie zwei dedizierte Netzwerk-Adapter zur externen Storage-Anbindung. Diese müssen jeweils über zwei Schnittstellen verfügen (QSFP112), Dual-Port) und wahlweise als 400Gb/s NDR InfiniBand / 400Gb/s oder 200Gb/s Ethernet konfigurierbar sein. Weiterhin verfügt der Server über zwei Management-Netzwerkadapter (1GbE NIC mit RJ45-Anschluss; 1GbE Host Baseboard Management Controller (BMC) mit RJ45-Anschluss).

Für das Betriebssystem stehen 2 x 1.9 TB M.2 NVMe SSDs zur Verfügung; für lokalen Storage stehen 8 x 3.84 TB zur Verfügung, die aus Performancegründen mit NVMe Modulen realisiert werden müssen.

Der Server ist luftgekühlt (Strömungsrichtung von vorne nach hinten), in 19-Zoll Bauweise ausgeführt und benötigt nicht mehr als 10 Höheneinheiten im Rack. Die Stromversorgung ist auf 230V AC / 50 Hz ausgelegt. Es werden Anschlusskabel mitgeliefert, mit denen der Server an bereitgestellte C19/C20 Rack-PDUs angeschlossen werden kann. Der Ausfall eines Netzteils kann toleriert werden (N+1 Redundanz).

Der Server wird komplett betriebsfertig konfiguriert geliefert (NVIDIA DGX OS); es ist keine weitere Installation von Software notwendig.

Zusätzlich zu dem Kauf der benannten Lieferleistung werden auch **Service- und Supportleistungen des Herstellers NVIDIA im Rahmen des „NVIDIA Enterprise Business Standard Supports“** für eine Laufzeit von 3 Jahren eingekauft (s. Anhang 3).

Damit diese Dienstleistungen auf entsprechend qualifiziertem Niveau erfolgen, muss der Lieferant einen „Elite Partner Status“ von NVIDIA haben. Erfolgt eine solche Beauftragung, verpflichtet sich der Auftragnehmer zu dieser entsprechenden Leistungserbringung.

2.2 Lieferung von Hardware

Die Mindestanforderungen, welche an den Server gestellt werden, sind im Dokument „**Mindestanforderungen Serverkomponenten**“ hinterlegt und mit „A“ (A-Kriterium = Ausschlusskriterium) gekennzeichnet ersichtlich. Die Nichterfüllung der als A-Kriterien kenntlich gemachten Mindestbedingungen führt zwingend zum Ausschluss des Bieters vom weiteren Verlauf des Vergabeverfahrens.

2.3 Serviceleistungen

2.3.1 Anlieferung

Ziel: Die Anlieferung der Produkte für das UKE erfolgt an die vereinbarte Verwendungsstelle

Die Anlieferung der Produkte für das UKE soll an die Verwendungsstelle in der Martinistr. 52, 20246 Hamburg, Gebäude N25, "Colocationspace" (00.2.015) erfolgen. Der Lieferant hat sicherzustellen, dass die vom UKE bestellten Produkte an die beidseitig vereinbarte Verwendungsstelle geliefert werden. Für die Installation ist ein Server-Lift erforderlich und es sind zwei Stufen zum Serverraum zu überwinden. Vor der Lieferung ist seitens Auftragnehmer in enger Abstimmung mit dem Auftraggeber sicherzustellen, dass sämtliche benötigten Hilfsmittel entweder seitens des Auftraggebers gestellt werden können oder durch den Auftragnehmer mitzubringen sind. Im Vorhinein der Anlieferung hat eine Abstimmung mit dem Klinikum zu erfolgen, wann und in welchem Zeitfenster konkret die Ware geliefert wird. Gemäß dieser getroffenen Absprache hat die Anlieferung dann zu erfolgen.

2.3.2 Aufbau und Installation

Ziel: Der Server wird im Rahmen der Anlieferung durch den Auftragnehmer betriebsbereit installiert, übergeben und eingewiesen

Der GPU-Server ist im Rahmen der Anlieferung vor Ort durch den Lieferanten im Server-Rack zu verbauen und hinsichtlich Stromversorgung und Netzwerkanbindung zu installieren. Ebenfalls ist die erforderliche Software auf Betriebssystemebene, sowie zur GPU-Interkonnektivität vorzuinstallieren bzw. vor Ort zu installieren und gemäß den Gegebenheiten im UKE zu installieren. Das Deep-Learning-Team am UKE ist entsprechend in die Verwendung des Systems einzuweisen.

2.4 Gewährleistung und Garantie

Alle im Rahmen des Vertragsgegenstandes gelieferten Komponenten müssen mangelfrei funktionieren. Detaillierte Regelungen zur Gewährleistung finden sich im EVB-IT Systemlieferungsvertrag. Der Lieferant

wird etwaige Herstellergarantien an das UKE weiterreichen. Service- und Supportleistungen des Herstellers NVIDIA

Die vollständigen Support-, Wartungs- und Instandhaltungsleistungen in Bezug auf sämtliche Hard- und Softwarekomponenten werden durch den Hersteller NVIDIA im Rahmen des „NVIDIA Enterprise Business Standard Support“ über eine Laufzeit von 3 Jahren nach betriebsbereiter Zurverfügungstellung und Abnahme durch das UKE erbracht, was dem UKE bei Auftragsvergabe durch den Auftragnehmer nachgewiesen werden muss. Die Supportleistungen müssen denjenigen im „**NVIDIA DGX Enterprise Support Data Sheet**“ (Anhang 3) und den Dokumenten **Mindestanforderungen Serverkomponenten** (Anhang 2) entsprechen.

2.5 Verbleib aller Datenspeicher beim Auftraggeber

Ziel: Erhalt des Eigentums an jeglicher Art von Datenspeichern bei deren Ersatz

Sollte ein in der Hardware integrierter Datenspeicher in Folge eines Mangels ausgebaut und ersetzt werden, verbleibt das Eigentum an diesem, beim Auftraggeber, um diesen autonom entsorgen zu können ("Media Retention").

2.6 Entsorgung

Ziel: Die Entsorgung der Altsysteme durch den Auftragnehmer soll sichergestellt werden.

Die Altsysteme – abgesehen der Datenspeicher (vgl. 2.6) – werden nach Aufforderung durch den Auftraggeber vom Auftragnehmer zurückgenommen.

Das UKE legt Wert auf eine umweltgerechte Entsorgung sowie fachgerechte Vernichtung der Altsysteme. Der Auftragnehmer hat daher sicherzustellen, dass die Entsorgung entsprechend der gesetzlich vorgesehenen Regelungen des ElektroG (insbesondere § 10 II, § 11 und § 12 ElektroG) vorgenommen wird. Grundsätzlich verbleiben die Datenspeicher der Altsysteme beim UKE.

Es wird ein Vernichtungsprotokoll der Altsysteme in die CMDB integriert.

3 Anhänge

3.1 Anhang 1: Serverkomponenten

Die folgende Tabelle beschreibt die Spezifikation des Deep-Learning-Servers. Der Server ist ausschließlich als 10 HE, 19" Rackserver anzubieten.

Grafikprozessor (GPU)	8 x NVIDIA Blackwell B300 GPUs, PCIe 5.0, 288 GB VRAM pro GPU, 2304 GB VRAM insgesamt mit der folgenden Performance-Spezifikation: – 72 petaFLOPS im Training – 144 petaFLOPS in der Inference – Unterstützung von FP64, TF32, BF16, FP16, FP8, FP5, FP4, INT8 Datentypen
GPU-Interkonnektivität	NVLink Switch mit 1800 GB/s bidirektionaler Bandbreite 14.4 TB/s aggregierte NVLink-Bandbreite
CPU	Dual Intel® Xeon® Platinum 6776P Processor, 128 Kerne insgesamt, 2.30 GHz (Basisfrequenz), 3.90 GHz (max. Boost)
Arbeitsspeicher	2 TB DDR5 RAM, erweiterbar auf 4 TB DDR5 RAM, ohne, dass die bereits verbauten DDR5-Module ersetzt werden müssten
Netzwerk	8x OSFP ports serving 8x single-port NVIDIA ConnectX-8 Up to 800Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU
Systemspeicher	Betriebssystem: 2x 1.9TB NVMe M.2 Speicher: 8x 3.84 NVMe U.2
Fernwartung	via Baseboard Management Controller (BMC)
Software	Vorkonfiguriertes Deep Learning optimiertes Betriebssystem (NVIDIA DGX OS).

3.2 Anhang 3 NVIDIA DGX Systems Enterprise Services



NVIDIA DGX™ Systems Enterprise Services provide you with comprehensive system support and access to services for both your NVIDIA DGX hardware and software (OS and firmware). With expansive levels of support and value-added services, NVIDIA provides you with the expertise and proven methodologies when and where you need them.

As AI workloads move from the fringe to the center of business, the need for defined service and support becomes necessary. NVIDIA brings to market the most advanced and robust experience across industries.

- > **Expert support:** To ensure business continuity, NVIDIA provides various levels of support, and has a deep connection with the product and engineering teams to keep DGX systems operating at peak performance.
- > **Faster ROI:** For faster results, NVIDIA offers onboarding, optimization, training, and certification services for employees and teams throughout the organization.
- > **Proactive and customized guidance:** For proactive recommendations for your data science practice and data center operations, NVIDIA experts provide personalized instructions and guidance.

For more information visit the [Enterprise Support and Services User Guide](#).

NVIDIA Enterprise Business Standard Support

All initial purchases of NVIDIA DGX systems include a three-year Enterprise Business Standard Support service for both hardware and software (OS and firmware) stacks. Initial four-year and five-year contracts are also available. Contracts can be renewed upon expiration. Below are the details of what is included with the Enterprise Business Standard Support service.

Entitlement	NVIDIA Enterprise Business Standard Support
Support Availability	> Cases accepted via the support portal for 24/7 support > Escalation support during a customer's local business hours (9 a.m. – 5 p.m, Monday–Friday) > Support provided by NVIDIA experts and engineers for timely resolution > Enhanced partner interaction on issues
Onsite Hardware Support	> Onsite engineer to replace field replaceable units (FRUs) > Advanced return material authorization (ARMA) for next-business-day delivery¹
Remote Support	> Remote hardware and software support
Software Updates and Upgrades	> Access to the latest NVIDIA DGX software (OS and firmware)
NVIDIA DGXperts and Curated Content Access	> Quick answers on AI problems from NVIDIA DGXperts² > Access to webinars, special events, and other curated content
Case Logging	> Comprehensive enterprise support > Enterprise Support Portal

Entitlement	NVIDIA Enterprise Business Standard Support
Portal Access Benefits	Full access to a knowledge base via the NVIDIA Enterprise Support portal, including: ➤ Insight into which knowledge base articles, tips, and tutorials are trending and have proven most valuable ➤ 24/7 access to NVIDIA Enterprise Support portal ➤ Create and manage support cases ➤ Manage entitlements ➤ Access to critical and general announcements ➤ Download software
Service Terms	➤ Three years included with initial purchase ➤ Extended service options available beyond three years

Beyond Standard Support

NVIDIA offers a wide selection of fee-based services and support for teams using NVIDIA DGX systems that can be purchased at any time. These services are only eligible for solutions with an active NVIDIA Enterprise Support contract.

Extended Support

Service	Description
Business Critical	All support referenced in NVIDIA Enterprise: ➤ Business Standard Support ➤ Live NVIDIA agent access 24/7 ➤ One-hour initial response for Severity 1 issues 24/7 ➤ Two-hour initial response for Severity 2 issues 24/7 ➤ Labor onsite for FRUs

Value-Add Support Services

Service	Description
<u>Technical Account Manager (TAM)</u>	Access to an NVIDIA service relationship manager, a TAM, who works as a liaison between customers and NVIDIA's elite technical resources. A standard TAM manages a single NVIDIA solution.
<u>Premier Technical Account Manager (PTAM)</u>	Access to an NVIDIA service relationship manager, a PTAM, who works as a liaison between customers and NVIDIA's elite technical resources. PTAMs support and manage two or more NVIDIA product families.
<u>Media Retention Services</u>	Allows customers with sensitive data protection requirements to retain products during an RMA action.
<u>Onsite Spares Service Program</u>	Allows customers to purchase select spares to augment their standard service agreement.

NVIDIA Infrastructure Services (NVIS)

Service	Description
Implementation	Custom installation and deployment of the DGX platform with expanded knowledge of NVIDIA networking solutions.

Education Services

Service	Description
<u>AI Infrastructure and Operations Fundamentals</u>	A self-paced course designed for enterprise IT professionals and administrators.
<u>NVIDIA-Certified Associate - AI Infrastructure and Operations</u>	This certification is an entry-level credential that validates the foundational concepts of AI computing related to infrastructure and operations.

Administration Courses

An instructor-led course designed to help customers administer their DGX systems. Available in private or public settings.

- [DGX SuperPOD™ Administration](#)
 - [DGX BasePOD™ Administration](#)
 - [DGX Systems Administration: private workshop or public bootcamp](#)
 - [NVIDIA AI Enterprise Administration: public training](#)
 - [Base Command Manager Administration: private workshop](#)
-

Ready to Get Started?

For information about these services, please visit the
DGX Support page at:
nvidia.com/en-us/data-center/dgx-systems/support/

Or contact enterpriseservices@nvidia.com

To renew your DGX Support contact
RenewalSales@NVIDIA.com

1 Next business day service may not be available in all regions. Please contact your local sales representative for availability.

2 NVIDIA DGXperts are AI-fluent practitioners composed of solutions architects, engineers, and the wider NVIDIA organization, who have built a wealth of experience over the last decade to offer prescriptive guidance and advice to help DGX customers fast-track AI transformation.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, DGX, DGX BasePOD, and DGX SuperPOD are trademarks and/or registered trademarks of NVIDIA Corporation. All company and product names are trademarks or registered trademarks of the respective owners with which they are associated. Features, pricing, availability, and specifications are all subject to change without notice. 4898800. MAR26

